DOCUMENT RESUME

ED 455 640 EC 308 528

AUTHOR	Miller, Phyllis, Ed.
TITLE	Testing: Let's Put It to the Test.
INSTITUTION	American Mensa Education and Research Foundation, Arlington, TX.
PUB DATE	2000-00-00
NOTE	152p.; Theme issue. Published three times a year.
AVAILABLE FROM	Mensa Education and Research Foundation, 1229 Corporate Dr. West, Arlington, TX 76006-6103. Tel: 973-655-4225; Fax: 973-655-7382; e-mail: millerp@mail.montclair.edu.
PUB TYPE	Collected Works - Serials (022)
JOURNAL CIT	Mensa Research Journal; n45 Fall 2000
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Adolescents; Adults; Creative Thinking; *Creativity; *Cultural Differences; Divergent Thinking; Evaluation Methods; Females; *Gifted; *Intelligence Tests; Psychopathology; *Self Evaluation (Individuals); *Test Validity

ABSTRACT

In this journal issue, articles examine various aspects of testing intelligence, creativity, and psychopathology. Featured articles include: (1) "Monglottosis: What's Wrong with the Idea of IQ Meritocracy and Its Racy Cousins?" (Johan W. Oller, Jr.), which shows empirically and theoretically that even nonverbal IQ tests mainly measure powers of reasoning accessed through the primary language of the test-takers and that verbal IQ scores assess proficiency in the language of the tests; (2) "Are Americans Becoming More or Less Alike? Trends in Race, Class and Ability Differences in Intelligence" (Wendy M. Williams and Stephen J. Ceci), which discusses findings indicating there is no evidence that dysgenic trends have caused declining American students' test scores and that there is a growing convergence across racial, socioeconomic, and ability related segments of American society; (3) "Self-Report Measures of Intelligence: Are They Useful as Proxy IQ Tests?" (Delray L. Paulhus and others), which discusses findings that indicate use of indirect and direct measures failed to yield valid results; (4) "Gifted--Through Whose Cultural Lens? An Application of Postpositivistic Mode of Inquiry" (Jean Sunde Peterson), which discusses findings that indicate the model of inquiry can be useful for those who seek new ways to conceptualize giftedness; (5) "Is the Proof in the Pudding? Reanalysis of Torrance's (1958 to Present) Longitudinal Data" (Jonathan A. Plucker), which discusses findings that just under half of the variance in adult creative achievement could be explained by divergent thinking test scores, with the contribution of divergent thinking being more than three times that of intelligence quotients; and (6) "Rorschach Interpretation with High-Ability Adolescent Females: Psychopathology or Creative Thinking?" (Kristen W. Franklin and Dewey G. Cornell), which discusses findings that higher scores on the Rorschach Schizophrenia Index among gifted female adolescents were correlated with healthy emotional adjustment. (Articles include references.) (CR)

ED 455 640

Testing: Let's Put It to the Test.

Phyllis Miller, Editor
Mensa Research Journal 45
Fall 2000

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Miller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

EC 308 528

BEST COPY AVAILABLE

BEST COPY AVAILABLE

Published by Mensa Education and Research Foundation and Mensa International, Ltd. ™

MENSA Research Journal 45

esting: Let's Put It to the Test

Fall 2000

BEST COPY AVAILABLE

EC 308528

MENSA Research Journal 45

Fall 2000

Table of Contents

The Mensa Education and Research Foundation 3

Editor's Preface 5

Notes, Quotes and Anecdotes
by Francis Cartier, associate editor 6

Monoglottosis: What's Wrong with the Idea of IQ Meritocracy and Its Racy Cousins?
by John W. Oller, Jr. 10

Are Americans Becoming More or Less Alike? Trends in Race, Class and Ability Differences in Intelligence
by Wendy M. Williams and Stephen J. Ceci .49

Self-report Measures of Intelligence: Are They Useful as Proxy IQ Tests?
by Delroy L. Paulhus, Daria C. Lysy and Michelle S.M. Yik 69

Gifted — Through Whose Cultural Lens? An Application of Postpositivistic Mode of Inquiry
by Jean Sunde Peterson 94

Is the Proof in the Pudding? Reanalysis of Torrance's (1958 to Present) Longitudinal Data
by Jonathan A. Plucker 117

Rorschach Interpretation with High-ability Adolescent Females: Psychopathology or Creative Thinking?
by Kristin W. Franklin and Dewey G. Cornell 135

© 2000 Mensa Education and Research Foundation

Mensa Research Journal is published three times a year by the Mensa Education and Research Foundation, Dr. Michael Jacobson, president, 1840 N. Oak Park Ave., Chicago, IL 60635-3314.

Staff:

Editor • Phyllis Miller, 23 Lexington Road, Somerset, NJ 08873, MRJ@merf.us.mensa.org

Associate Editor • Francis Cartier, 1029 Forest Ave., Pacific Grove, CA 93950, fcar889755@cs.com

Assistant Editor • Al Derr, 831 Lombardy Drive, Lansdale, PA 19446, derr@vu-vsi.ee.vill.edu

Foundation Assistant • Patty Wood, Mensa Education and Research Foundation, 1229 Corporate Drive West, Arlington, TX 76006-6103, (817) 607-0060 ext. 111

Mensa Research Journal
Editorial Advisory Board

Francis Cartier, Ph.D.
Dorothy Field, Ph.D.
Annette Greenland, Ph.D.
Ilene Hartman-Abramson, Ph.D.
Edward J. Haupt, Ph.D.
Michael H. Jacobson, Ph.D.
Ken Martin, Ph.D.
Caroline Mossip, Psy.D.
Charles A. Rawlings, Ph.D.
Abbie F. Salny, Ed.D.
Hirsch Lazaar Silverman, Ph.D.

The *Journal* is published three times a year. Subscribers will receive all issues for which they have paid, even if frequency of publication varies. Membership in Mensa is not required. To subscribe, order back issues, or report address changes, write to *Mensa Research Journal*, 1229 Corporate Drive West, Arlington, TX 76006-6103.

The Mensa Education and Research Foundation

Mensa, the high IQ society, provides a meeting of the minds for people who score in the top 2 percent on standardized IQ tests. As an international organization with thousands of members worldwide, Mensa seeks to identify and foster human intelligence; encourage research in the nature, characteristics and uses of intelligence; and provide a stimulating intellectual and social environment for its members.

The first two of these purposes are largely carried out by the Mensa Education and Research Foundation (MERF). MERF is a philanthropic, nonprofit, tax-exempt organization funded primarily by gifts from Mensa members and others. MERF awards scholarships and research prizes, publishes the *Mensa Research Journal,* and funds other projects consistent with its mission.

For more information about the Mensa Education and Research Foundation, write to MERF, 1229 Corporate Drive West, Arlington, TX 76006-6103. For information about joining American Mensa, call 1-800-66-MENSA. See also www.us.mensa.org. If you reside outside the U.S., see the Mensa International Ltd. Web site at www.mensa.org.

To renew your subscription or subscribe to the *Mensa Research Journal*, just fill out the form on the next page (or a photocopy of it) and mail it with a check in the appropriate amount. Your present subscription expires after the issue number you will find on your mailing label. Look at it now.

To: MERF, 1229 Corporate Drive West, Arlington, Texas 76006-6103

Please enter my subscription to the *Mensa Research Journal*.

☐ New ☐ 3 issues U.S. $21 (outside U.S. $25)
☐ Renewal ☐ 6 issues U.S. $42 (outside U.S. $50)
☐ Sample *Mensa Research Journal* U.S. $7 (outside U.S. $9)

Subscription amount enclosed $ ______________________
Please make payments in U.S. funds.

I would like to make the following tax-deductible contribution to MERF.

☐ Contributor ($25 to $99)	$ __________	Allocation (if desired)
☐ MERF Donor ($100-249)	$ __________	☐ General support
☐ Bronze Donor ($250-499)	$ __________	☐ Scholarships
☐ Silver Donor ($500-999)	$ __________	☐ Gifted Children Programs
☐ Gold Donor ($1,000-2,499)	$ __________	☐ Addition to endowed fund
☐ Platinum Donor ($2,500-)	$ __________	named __________

I would like my contribution to be a ☐ Memorial honoring ______________
☐ Tribute honoring ______________

Please send acknowledgement or notice to ______________________
Address ______________________
City/State/Zip ______________________

Name and address of contributor/subscriber:

Name ______________________ Member # __________
Address ______________________
City/State/Zip ______________________ MRJ

For Office Use
Date received: ________ Total $ ________ Bank: ________ Date entered: ________
Sub code: ________ First issue number: ________ Date sent: ________

Editor's Preface

If you're in Mensa, you once took a test. Maybe it was the GRE or the SAT, maybe the Army test, maybe the Mensa entrance exam. Whatever it was, you did well on it, well enough to be in the top 2 percent. And you probably felt very proud of yourself.

But did you ever wonder exactly what was being measured when you took the test? And who was doing the measuring? Since IQ tests were first developed, there have been many variations and refinements, most of them in response to changes in society over the years. Today we worry about culture-free or culture-fair, we worry about test-takers with disabilities, we worry about tester bias, so whatever it is we are measuring today is different from what we measured years ago, maybe even different from when you took the test.

The field of measurement is one that has expanded mightily since M. Binet set pencil to paper. In this issue of the *Mensa Research Journal,* a number of researchers examine various aspects of testing intelligence, creativity and psychopathology. Some of their conclusions are disturbing — sort of, why are we doing this anyway — but all are important if we as a society continue to use various types of tests to determine the outcome of people's lives: where they go to school, what occupations they may choose, what organizations they can join, and even if they're nuts.

So, read on, MacDuff, and when you're done, think about the test you took that got you into Mensa.

Phyllis Miller
Editor

Notes, Quotes, and Anecdotes

It's probably wise to remind you that the opinions expressed in the *Mensa Research Journal* are strictly those of the authors. In fact, they are the opinions of the authors at the time they were written. Neither American Mensa nor MERF holds any position on the contents of the *MRJ*.

• John Oller's article in this issue should raise some controversy. That would be nothing new for John, who has been introducing innovative ideas into testing for as long as I've known him. We first met when we had a mutual interest in testing proficiency in English as a foreign language about 30 years ago. I haven't always agreed with him on language testing and, in fact, we've had considerable correspondence on his present topic. Nevertheless, I've always admired how John's mind works; it looks meticulously at things most others don't see at all.

• Which reminds me of another brilliant non-conformist mind I have long admired: that of Lee Thayer. I met him first, I think, back in the early 1950s, when I was editor of *The Journal of Communication (J. of C.)* published by the National Society for the Study of Communication (later infelicitously renamed the International Communication Association). I was president of NSSC in 1959 and Lee was president in 1968. This afternoon, I was rereading his still-thought-provoking article, "On 'Doing' Research and 'Explaining' Things," in the *J. of C.*, summer 1983. Ponder these quotations. Lee is writing about research in communication, but it applies to research in intelligence, too.

> Most of what we call "research"... is done first in order to get a degree — to be properly credentialed for employment in the academic-research establishment; second, for purposes of promotion and tenure — to secure a sinecure; and third, to gain status and prestige in one's field. One accrues and displays the accoutrements of seniority and success in the academic-research establishment in much the same way as those of power and privilege do in the military-industrial complex.

Lee doesn't exactly say that the topics and methods of research are inevitably determined by these factors, but he was concerned about their influence. He goes on to remark that:

> The only part of the massive and growing product of the academic-research establishment that [captures] public attention is that which captures the attention of the media mongers — the pop, the cultish and faddish, the topical, the esoteric, the offbeat, the off-color, the dramatically problematic, the popularized and the populizable — translated and repackaged to make it seem more relevant, and much more certain than it was or is.

Does that ring a bell for you about media reports of research into intelligence?

- The homepage for the Center for the Neural Basis of Cognition lists and connects to virtually every source of information on how the brain works. Check it out at http://cnbc.cmu.edu/other/homepages.html. Another interesting site is the Los Alamos National Laboratory's "This is Mega-Mathematics." Go to www.c3.lanl.gov/mega-math for links to several other math and analytical sites.
- Lauren Resnick, University of Pittsburgh, had an article in the March 1999 *APA Newsletter for Educational Psychologists* with the intriguing title, "From Aptitude to Effort: Learnable Intelligence and the Design of Schooling." After tracing and criticizing the theories that have dominated our schools for many years (such as the archaic but still pervasive practice of grading achievement on the bell curve), Resnick says, "If American schools are ever to break out of their aptitude-centered and drill-and-practice traditions, something new is needed." Resnick "...challenges the notion that education must choose between [Thorndike's] passive drill pedagogies and [the 'progressive schools'] child-centered discovery pedagogies. Consider the 'reading wars' as an example." The unproductive debate is usually over the supposedly bi-polar issue of

> "whether children should first be systematically taught the print-to-sound code (the 'phonics' approach) or first be immersed in a rich environment of books and writing and allowed to induce the code over time (the 'whole language' approach). The popular image of the phonics approach consists of Thorndike-like drills, with every sound-spelling pattern being taught and practiced. In contrast, the popular image of whole language equates it with a radical child-centered approach to teaching reading, in which children pursue their own choices and learning the print-sound code is left to chance."

Research has shown, however, that both theories are only partly true. "[S]ome form of phonics instruction is necessary [and] learning to comprehend what one reads needs active and explicit attention as well."

Resnick blames some school administrators and teachers [I'd add politicians and the media] who seem committed to oversimplifying the process. However, that's not the most important thing Resnick has to say about the current state of education.

"People differ widely in their beliefs about effort in learning or problem-solving situations," she says.

There are performance-oriented people and learning-oriented people. The former tend to view ability as unchangeable; the latter strive to increase ability. "People with learning goals especially direct their efforts to strategic problem solving and learning. These are the very activities that cognitive research tells us can create aptitude."

We need a new way, which she calls "knowledge-based constructivism," that honors each student's right to expert instruction, regardless of his or her initial ability. That will require appropriate training of both teachers and school administrators.

You and I will surely agree with Resnick but still wonder how that might come about. We might also wonder whether the theory of constructivism has sufficiently come of age to be a viable basis for instructional methodology.

- Richard M. Ingersoll explores what may be the most critical stumbling block to the appropriate teacher training that Resnick hopes for in his well-documented article, "The problem of underqualified teachers in American schools," *Educational Researcher*, March 1999, pp. 26-37. "One of the least recognized causes of [poor quality teaching is known as] out-of-field teaching — teachers assigned [by their principals] to teach subjects for which they have little training or education." James Conant called attention to it in 1963. In 1985, Albert Shanker called it education's "dirty little secret."

Ingersoll admits it's hard to measure out-of-field teaching, so he tried to find out "how many of those teaching core academic subjects at the secondary level do not have even minimal credentials — neither a major nor a minor — in their teaching fields [or a related subject]." Answers: one-third of the math teachers, about one-quarter of the English teachers, about one-fifth of the science teachers, and about one-fifth of social studies teachers. The numbers without teaching credentials in the subjects they are teaching are similar and have changed little from the late 1980s to the mid-1990s. Of course, he points out, some teachers may have made themselves competent in the out-of-field subject and often also teach classes for which they are fully qualified. However, Ingersoll's study showed that "out-of-field" assignments are associated with decreases in teachers' morale and commitment. Moreover, one might also ask, does out-of-field teaching have any effect on the legitimacy and authority of teachers and hence, on classroom discipline?

Ingersoll spends several pages on why out-of-field teaching happens, dispelling some of the usual excuses. He doesn't touch on underqualified teachers of the gifted or otherwise exceptional students, but the implications are fairly clear.

- *Time*, July 5, 1999, reported research at the University of California at Irvine suggesting that 10 minutes of Mozart's *Sonata for Two Pianos in D major (K. 448)* can increase students' scores on that part of an IQ test related to spatial-temporal reasoning, which is important to math skills. Of course, when the rest of the press picked up the story, the Irvine research was irresponsibly expanded to imply that classical music played to babies and infants would increase their intelligence, which isn't at all what the Irvine researchers said. Furthermore, other researchers have been unable to replicate even the limited Irvine results, cf. two research reports *Psychological Science*, July 1999. But hey! I know it works because my two kids were raised in an environment of nearly all-day classical music and grew up gifted! Well, anyway, they grew up appreciating classical music. It worked for young Mozart, too. Aren't real anecdotes more valid than scientific research?

- My partly illegible old notes of a talk by Keith Simonton a couple of years ago (at an APA convention?) just surfaced on my cluttered desk. I believe my notes say that he said that "a creative thought" is recognition of a new analogy when meditation permits "remote ideas to bump into each other." That's close to the position I stated some years ago that every new idea derives from discovery of a previously unrecognized relevance of two older ideas. Except that I eschew the term meditation; I prefer concentration. I also disagree with Simonton that "creativity is unteachable," if indeed that's what he said.

• *Dumb things smart people do.* The Seiko Company has come up with a handheld "Quicktionary Reading Pen with Speech" designed "to help persons who are learning to read English." If you stroke it across a word you don't know, it pronounces it aloud and, from its small screen, you can read the definition. Brilliant bit of technology, huh? Well, maybe. The ad shows it scanning the word *quick*. The definition it gives is:

Having or functioning rapidly and energetically.

Sounds right? Now *think*. How many students of English who are at the level to need a definition of *quick* are going to know the words *functioning, rapidly* or *energetically*? And as for the grammar of that definition, well . . . .

Francis Cartier
Associate Editor

Monoglottosis: What's Wrong with the Idea of the IQ Meritocracy and Its Racy Cousins?

John W. Oller, Jr., University of Southwestern Louisiana

For about 100 years, certain enthusiasts have claimed that IQ tests measure innate intelligence (Binet and Simon 1905; Brigham 1923; Eysenck 1971; Herrnstein 1973; Herrnstein and Murray 1994; Jensen 1969, 1980, 1984, 1995; Lynn 1978, 1979) and show racial differences. These ideas have roots in "social Darwinism" (Darwin 1874) and the eugenics movement (Galton 1869) — the aim to "purify" the gene pool. Linked to these racy theories are the over-representation of minority language children in classes for the mentally retarded, language disordered, etc., and their under-representation in classes for the gifted (Ortiz and Yates 1983; Oakland and Parmelee 1985). In opposition to the IQ elitists, others have claimed that the data are misconstrued (Figueroa 1989; Fraser 1995; Gould 1981, 1995; Isham and Kamin 1993; Jacoby and Glauberman 1995; Macnamara 1966, 1972; Mercer 1973, 1984; Valdes and Figueroa 1994) or perhaps irrelevant (Gardner 1983, 1993, 1995). While the research of Herrnstein, Murray, and Jensen (not to mention Carroll 1993, 1995; Sternberg 1996; and others) cannot be ignored, it can be shown that the IQ enthusiasts have largely discounted acquired language/dialect proficiency as a factor in their tests (Oller and Perkins 1978; Oller 1978; Oller, Chesarek, and Scott 1991; Oller and Jonz 1994). Monoglottosis, near total language/dialect blindness, is partly to blame. This condition accounts for Herrnstein's "meritocracy" theory that intellectual cream rises to the top. But do IQ tests measure "innate" intelligence? It is shown here empirically and theoretically that even "nonverbal" IQ tests mainly measure powers of reasoning accessed through the primary language of the test-takers and that "verbal" IQ scores assess proficiency in the language of the tests. The IQ literature needs to be reconceptualized.

Introduction

Why is it that IQ testers have traditionally supposed their tests are capable of seeing right through language dialect differences to the genetic level of intelligence? Even Binet and Simon, the creators of the first IQ tests, not to mention today's staunchest defenders of IQ testing, Jensen, Herrnstein, Lynn, Eysenck, Gordon (see the reference list), and the few others of whom these enthusiasts have approved, embraced the claim that IQ tests measure *innate ability independently of the language/dialect and experience of the test-takers.* I call this the innateness premise. Here I will endeavor to show that the original creators of IQ tests, along with a powerful vocal minority of present-day measurement

---

Reprinted from *Applied Linguistics,* vol. 18, no. 4, 1997.

specialists, still hold tenaciously to this innateness premise.[1] They insist that observed differences in IQ scores are linked to race (that whites are smarter than blacks) and that the more intelligent rise to the top. The latter idea is what Herrnstein called the theory of the meritocracy (Herrnstein & Murray 1994: 511-12).

Why do these IQ enthusiasts believe that their tests are really measures of *innate* intelligence? It can easily be seen by anyone who examines the content of the tests, as we will in this paper, that all "verbal" IQ tests are utterly dependent on listening, speaking, reading, and writing in English or *some* other particular language. Verbal IQ tests *are* language/dialect tests, plain and simple. To defend against this obvious fact, the IQ enthusiasts protest that they do not rely on tests in English or any particular language exclusively, but that they also employ "nonverbal" or "performance" tests which allegedly do not require *any language at all.* They claim that these "culture-fair" and "language-less" "nonverbal IQ" tests are nearly pure measures of inborn intelligence (e.g. Jensen 1980: 646). Is this claim true? Are "nonverbal" IQ tests really free of dependence on any particular language or dialect?

Here I will endeavor to show that the claims generally made in defense of the innateness premise are not only unreasonable upon examination, but that they are demonstrably false. If my case is correct, IQ scores have quite generally been interpreted incorrectly and applied inappropriately, and have often resulted in needless harm to school children and others. To develop my argument, it will be necessary to establish what the views of the vanguard of IQ testers really are, where they came from, and how IQ tests have been applied and what they consist of. I aim to show why those views, uses, and interpretations are generally incorrect and how they can be set right.

Defining "Monoglottosis"

A key element in the whole discussion is the role played by different languages and dialects. There is an old riddle that asks, "What do you call a person who speaks two languages?" People usually say, "Uh,... bilingual." Next, the riddler asks, "And what do you call a person who speaks three languages?" The respondent nods confidently, "Trilingual." The riddler then follows with the clincher: "Then, what do you call a person who speaks only one language?" The respondent says, "Monolingual, right?" The riddler gives a gotcha-smile and says, "Nope. You call that person an American."[2]

---

[1] It is important to point out from the start that I am not claiming that the views of the elite vanguard of IQ testers are shared by the majority of psychologists and educators. On the contrary, the majority consensus generally disputes many of the racist claims of the highly vocal minority of IQ enthusiasts, see Sternberg (1985), Linn (1989), Carroll (1993), to name only a few of the dissenting majority. Nevertheless, I will show here that the vocal minority as documented widely by Hakuta (1986), Valdes and Figueroa (1994), Hamayan and Damico (1991), and others, has shaped opinions that have determined policy and practice in far too many cases.

The riddle is hyperbolic and just a little too hard on Americans because we have no exclusive hold on monolingualism (and many of us are not monolingual), but the riddle makes a valid point: *People who are monolingual (or monodialectal) are naturally language/dialect blind.* It is not at all obvious to them that they even require a language/dialect for thinking, conceptualizing, and expressing information.

Some time ago I proposed, tongue-in-cheek (Oller 1994), to call this special condition of language blindness *monoglottosis*. I had in mind the resemblance of this invented word to terms like "mononucleosis," "halitosis," and terminal "monolingualism." Just as a fish is normally unaware of the water it swims in, monolingual persons (and even some polyglots) are apt to remain unaware of the essential role language proficiency plays in their lives. Of course, monoglottosis does not affect one's ability to notice distinct forms of speech (as of the foreign-born) or signing (as of deaf communities). It is not a blindness to differences in speech (or what Saussure 1912 called *parole*). It does not prevent awareness of the surface-forms of languages because these are noticeable enough to anyone with almost any combination of the senses of sight, hearing, touch, taste, and smell. Rather, monoglottosis is a general unawareness of the languages or dialects that must be called on to make sense of the surface-forms of speech or other signs that enable communities to share abstract meanings (i.e., what Saussure 1912 called *langue*) — e.g., the conventional fact that the word "water" refers to water had to be discovered by Helen Keller (1908). Monoglottosis is a special blindness toward the general dependence of all sign-users on such conventions in some particular language/dialect.

What the IQ enthusiasts generally have done is to ignore — and in some cases to explicitly deny — that any particular language/dialect is needed to gain access to the abstract meanings that are manipulated in their IQ tests (Binet & Simon 1905; Brigham 1923; Gordon 1980; Jensen 1980, 1984; Herrnstein & Murray 1994). In this way, they demonstrate that they suffer from monoglottosis. In fact, the evidence shows (Cummins 1984; Figueroa 1989; Hakuta 1986; Hamayan & Damico 1991; Kamin 1995a; Oller & Perkins 1978, 1980; Oller 1978, 1979, 1983, 1992; Oller, Chesarek, & Scott 1991; Valdes & Figueroa 1994) that the proponents of racial differences in IQ scores and even some of their critics (e.g. Fraser 1995; Gartner, Greer, & Reissman 1974; Jacoby & Glauberman 1995; Mercer 1984; Sternberg 1996) are *remarkably impervious to the role that language proficiency plays in IQ tests, school achievement, and in any conceivable definition of the "meritocracy."*

Throughout the discussion to follow it is also important to keep in mind that IQ tests really have been used widely from the time of Binet and Simon (1905) forward to make judgments, *not about acquired language proficiency,* but about the *"native" intelligence of children and adults* (Brigham 1923; Figueroa 1989;

---

[2] This riddle was passed on to me by my brother, D.K. Oller.

Oakland & Parmelee 1985; Hakuta 1986; Valdes & Figueroa 1994; Yerkes 1921). Although the theoreticians may have often disagreed with prevailing trends (see note 1), the trends themselves have not been much affected by those disagreements. Many of the persons tested on English language tests incorrectly designated as measures of innate intelligence have been immigrants to the United States from non-English backgrounds (Gersten & Woodward 1994; Jitendra & Rohena-Diaz 1996; Hamayan & Damico 1991; Simich-Dudgeon 1992). Speakers of other languages in the schools have often been tested in English in spite of the criticisms and known deficiencies of the tests employed. As a result, *proficiency in the language of the tests (English in most cases) has generally been mistaken for native intelligence.* Of course, this error is not unique to the U.S., nor is it an error limited to testers who have used English.

The ubiquitous (but commonly denied) outcome of this erroneous interpretation of IQ test scores is that children and adults who were perfectly normal have often been misjudged to be mentally deficient and have been treated accordingly in the schools and workplace (Cummins 1984, 1986). To add insult to injury, as shown by Hakuta (1986; see also Lambert & Tucker 1972; Valdós & Figueroa 1994), educators and others suffering from monoglottosis have accepted the idea that IQ tests are measures of *innate* abilities and have gone on, in some cases (e.g., Gordon 1980; Jensen 1984), to argue that bilingualism itself was a *causal factor in producing the retardation or other alleged mental deficiencies* of persons from non-English speaking backgrounds. While IQ testers are by no means exclusively responsible for the common mislabeling of children of diverse language backgrounds, Ortiz and Yates (1983) documented that the over-representation of minority language children in classes for the educable mentally retarded persists — Hispanic children in Texas were more than 300 percent over-represented there quite recently. In a later study, Ortiz (1987) reviewed school records and concluded that "assessment procedures used to diagnose communication disorders virtually ignored the linguistically diverse nature of these learners" (Jitendra & Rohena-Diaz 1996: 42). Instead, traditional methods of assessing language proficiency in English were indiscriminately used by mostly monolingual teachers in order to assess learners from minority language backgrounds (Chamberlain & Medeiros-Landurand 1991; Hamayan & Damico 1991; Hayes-Brown 1984; Rivera & Simich 1981).

In 1991, Alan Ginsberg (see published papers in Simich-Dudgeon 1992) was asked why this sort of misdiagnosis was possible. "Why is it," Leonard Baca asked, "that minority language children are so commonly misdiagnosed as having special language impairments, communicative disorders, or other mental handicaps?" Ginsberg said that while "those problems may exist in *some* areas of the country, we don't really know why" (Ginsberg 1992). But Ginsberg's statement is simply not correct. We not only *know* that minority language children are commonly misdiagnosed in essentially *all* areas of the U.S., but we also know *why*. It is mainly because "verbal" IQ tests and related achievement tests and diagnostic procedures — i.e., tests dependent on English language proficiency — are routinely applied to so-called "bilingual"[3] persons who have not yet acquired English (Damico, Oller, & Storey 1983). This misguided practice is well-documented across the entire spectrum of education and is not limited to the diagnosis of mental disorders and learning disabilities. Figueroa (1989) has commented that "school psychologists are still testing bilingual students and those with limited English proficiency predominantly in English" (1989: 147 — also Damico, Smith, & Augustine 1996; Figueroa 1989; Jitendra & Rohena-Diaz 1996; Oakland & Parmelee 1985; Ortiz 1986, 1987; Ortiz & Polyzoi 1988; Rueda, Figueroa, Mercado, & Cardoza 1984; Simich-Dudgeon 1992).

Though IQ tests are not the only diagnostic tools that are misapplied with persons from non-English language backgrounds, they have been among the tests most frequently misused almost since the beginning of public debates about testing (as documented by Buros 1938-1972). In spite of all this, IQ tests are still the most vigorously defended of all existing mental measuring devices. Indeed, they are commonly held up (by a highly vocal minority of "experts") as being without bias and altogether above reproach (Gordon 1984; Herrnstein & Murray 1994; Jensen 1980, 1984, 1995). They are also the single most important diagnostic tool used in the *definition* of mental retardation and a host of other disorders and disabilities (Cummins 1984, 1986). Even Jensen (1980: 109) documents and subscribes to the fact that the *very definition of mental retardation* is explicitly linked with IQ scores by the American Association of Mental Deficiency (also see American Psychiatric Association 1980, 1987, 1994; World Health Organization 1993). Mercer (1984: 325) points out that the term "moron" was coined by Goddard (one of the early IQ testers involved in the creation of alpha and beta; see Yerkes 1921: 299) after the IQ movement and its accompanying eugenics program (see below) was well underway.

Jensen (1980) devotes a scant four and a half (out of 786 over-sized) pages to the effects of acquired proficiency in any particular language or dialect on IQ scores (1980: 603-607). To anyone suffering from monoglottosis, Jensen's remarks will seem perfectly sensible. But to show that they involve the special blindness imposed by that widespread syndrome, here is a quote with interpolations in square brackets to show where and how monoglottosis is affecting the argument:

> ...children from bilingual homes...generally perform better on standard tests [i.e. tests of IQ or achievement given in English] than do children from homes in which Spanish [or any language other than English!] is spoken exclusively...The language of the test (or of the E[xaminer]) makes less difference the longer S[ubject]s have attended English language schools, and the disparity

---

[3] Oddly, this term is often misapplied to every conceivable kind of non-English background. It has become a catch-all even for monolingual persons of any language background other than English wherever they happen to be found in English-speaking environments.

> between verbal [English language] and nonverbal test scores diminishes with increasing number of years in [an English speaking] school. (ibid.: 606)

Careful readers will note that when Jensen says "*verbal* test" he usually means "*English language* test" and when he says "in *school*" he usually means "in a *school where English is spoken*." If this were not so, his statements would make no sense. For instance, it is patently false to say that the disparity between verbal IQ scores in, say, Mandarin and nonverbal IQ scores will automatically diminish as a person is schooled in, say, French, or in fact in any language other than Mandarin. Also, there is no reason to suppose that scores in, say, Navajo will ever necessarily improve because the person spends more time in a school where, say, Spanish, or any language other than Navajo, is the language of instruction.[4]

If monoglottosis was a factor in Jensen's theorizing about intelligence (1969, 1980, 1984) it became even more so in Herrnstein and Murray (1994). They do not even mention the effect of language proficiency on any kind of "IQ" test, "verbal" or otherwise in 845 pages. No discussion of primary or secondary language proficiency will be found in their attractive volume, nor does it mention "language" or "dialect" in its index. However, there is no plausible basis for denying that variance in "verbal IQ" *must* consist largely of acquired proficiency in whatever language is used in those "verbal" tests (Oller & Perkins 1978; Gardner 1983, 1993, 1995; Carroll 1983, 1993; Figueroa 1989; Valdes & Figueroa 1994). In fact, once this idea was clearly stated, thoughtful psychometricians acknowledged that it had to be correct. This is because verbal IQ tests cannot fail to assess acquired language skills and still qualify as being "verbal" (Sternberg 1996). Even Jensen (1980) obliquely acknowledges this fact though he tries to argue that what "verbal" IQ tests *really* measure is innate intelligence because, he argues, the language part is easy (1980: 132). Following Gordon (1980), he even goes so far as to say that low verbal scores in English, for certain minority language groups, can be interpreted as showing that "the educational disadvantage of bilingualism may be largely the result of lower verbal aptitude *per se* than of a bilingual background" (Jensen 1984: 535). Jensen supposes that becoming "bilingual" is evidence of a low IQ to start with.

The role of acquired language/dialect proficiency

The vast literature on IQ testing shows that the vocal minority of IQ elitists are trapped in an invisible prison of cultural and linguistic myopia. As a result, they cannot appreciate the role that *acquired language proficiency* must and does play in so-called "IQ" tests. The phrase *acquired language proficiency* here means nothing more nor less than the ability to associate the sounds (or visible, kinetic, or otherwise noticeable signs), words, and sequences (the speech or other surface-forms) of any particular language (or any dialect), e.g., English, Zuni, American Sign Language, etc., with everyday experience. Since no infant is born a speaker, reader, writer, or signer of any particular one of the world's 5,000 or so languages (not counting distinct dialects; Grimes & Grimes 1993), it is clear that language proficiency is something that must be *acquired over time* (de Villiers & de Villiers 1979; Slobin 1987; Thelen & Smith 1994). The fruits of this process, *language acquisition,* can no more be directly inherited than the ability to ride a bicycle, read a book, or play an instrument can be passed genetically from parent to child.

---

[4] Jensen also seems to have fallen into another trap common to persons suffering from monoglottosis. He writes as if there were only one kind of multilingualism — the kind seen in speakers of Spanish and English. But, of course, in view of the fact that there are estimated to be about 5,000 languages in the world (Grimes & Grimes 1993), the kinds and degrees of multilingualism are not reducible to any particular case of bilingualism.

As Wing (1981: 40) put it:

> The child develops theories concerning the relationships among people, objects, and events, which are then expanded, modified, or rejected in the light of new information. This process is present from birth but becomes much more efficient with the development of symbolic language, which allows the labeling of concepts and their economic storage in the memory.

Thus, the child must gain access to *conventional* relations between surface forms in the target language (e.g., "dog," "bark," etc.) and contexts of experience ("there's a dog," "that dog is barking," etc.). Without access to the conventional applications of the linguistic signs in known contexts of experience particular languages/dialects cannot be acquired at all.

Surface-forms of language are not transparent

It is quite generally impossible from the surface-form of any linguistic sign, including any kind of complex syntactic structure, all by itself to guess just what its meaning is apt to be, as noted by Saussure (1912). We require access to the conventional relations between signs and at least some of their contexts of use. Which particular contexts may not matter much, but at least *some* contexts must be made available if anyone is to acquire any language. The problem for a dynamic theory of intelligence — in which a theory of language acquisition must be an important, if not the central, part — is to show precisely how a normal child can abstract signs from experience to form progressively more adequate sign systems (Piaget 1947; Oller 1996). As Carroll (1976: 27) pointed out, "Tests of 'intelligence' have always been the most prominent type of psychometric instrument. However great their interest in practical matters, all the leading figures in psychometrics — Binet, Spearman, Thurstone, and Guilford (to name but a few) —... have realized that to construct a theory of intelligence is to construct a theory of cognition." However, the IQ testers have generally left to others the task of showing the crucial role played by language signs in the mental development of human beings (e.g. Luria 1966; Luria & Yudovich 1959; Oller 1996; Peirce 1898; Piaget 1947; Vocate 1987; Vygotsky 1934).

A central question is how a suitably endowed organism can acquire any language. While genetic inheritance (as argued by Bickerton 1981, 1995; Chomsky 1965, 1995; Pinker 1994) is no doubt crucial to language acquisition, it does not supply enough information to enable anyone to develop proficiency in a given language without exposure to comprehensible uses of that language in ordinary contexts of experience. In fact, without exposure to noticeable links between the surface-forms of any target language and the ordinary contexts of experience, acquisition is demonstrably impossible. A strict demonstration of this claim has been worked out (Oller 1993, 1995, 1996) following Peirce's (1865, 1898) method of "exact logic." At every step of the argument the necessity for that particular step is demonstrated *before* it is taken. This is done by showing for each step of the argument that unless it is taken, an inevitable contradiction must arise. Omitting certain elaborations given elsewhere (Oller 1996), the theory shows (unsurprisingly, but rigorously) that language acquisition requires access to ordinary contexts where the language in question is used according to its own conventional requirements or grammar. The upshot of that argument (again, unsurprisingly) is that *the language of IQ tests must be acquired and not innate.*

Monoglottosis in the History of IQ Tests

Nevertheless, the most vocal IQ enthusiasts began with and have persisted in the notion that every genuine IQ test is mainly a measure of *inborn, innate, genetically-determined intelligence.* This has been true since the creation of what Robertson (1972) called the "granddaddy" of all IQ tests — the Binet-Simon IQ battery (transformed in 1916 by Terman and Goddard into the Stanford-Binet). This was stated plainly by Binet and Simon (1905) and has been maintained ever since by enthusiasts such as Brigham (1923: 97), Jensen (1980: 141ff.), Herrnstein and Murray (1994: 3), and the elite defenders of IQ testing. Binet and Simon (1905: 42) wrote:

> It is intelligence alone that we seek to measure, but disregarding insofar as possible, the degree of instruction which the subject [the child tested] possesses.... We believe that we have succeeded in completely disregarding the acquired information of the subject.

Binet would later retract this view (1911), but his second thoughts, and those of many other mainstream advocates of caution (e.g. Carroll, Gardner, Linn, Sternberg, etc.) would generally go unheeded even to the present day by the staunchest defenders of the innateness premise. Jensen has contended, further, that *innate intelligence* is essentially identical with the general factor that can be extracted by any number of analytical methods from almost any battery of mental tests. Spearman (1904) called that factor "general intelligence" and abbreviated it as "*g*." Jensen (1980: 224) said:

> We identify intelligence with *g*. To the extent that a test orders individuals on *g*, it can be said to be a test of intelligence. Not all so-called intelligence tests meet this criterion equally well,... Yet IQ tests such as the Stanford-Binet and the Wechsler Scales would probably correlate between .8 and .9 with a hypothetical true scale of *g* in the *normative* [my italics] population.

In all of this Jensen (1993, 1995) agrees with and approves of the views of Herrnstein and Murray (1994: 302-4) about the centrality of *g*. Setting to one side the often noted need to differentiate the factors other than *g* in mental tests, it remains curious that Jensen is obliged to refer to a "*normative* population" at all. In doing so, he inadvertently admits that IQ scores are not interpretable without reference to normative behaviors. Yet, if IQ tests really were direct measures of *g* — alias *innate intelligence* — the choice of a "normative population" would be quite irrelevant. The fact that such a choice is not irrelevant merely shows the inevitable truth that access to abstract meanings depends on linguistic *conventions, alias "norms of usage"* that differ from one language/dialect to another. Thus, Jensen inadvertently undermines his own case. IQ tests are perfectly universal neither in their conception nor in their application.

If languages were as transparent as the innateness premise pretends, the language used in IQ tests would have to be strictly "onomatopoetic" — i.e., the sounds or other signs would directly reveal their meanings. Natural languages most nearly achieve this kind of transparency in such sounds as "rat-a-tat-tat" imitating a machine gun, or "varoom, varoom" to show the double revving of an engine. For IQ tests to escape dependence on conventional linguistic norms, the sound sequences used in the instructions for each section and in every IQ item would have to provide direct access to meaning. In such a case, the tests would be perfectly onomatopoetic, and there would be no need for inventing "nonverbal" IQ tests with their dependence on "gesture and pantomime." In other words, the *conventional* and culturally unique basis of distinct linguistic systems is either ignored or explicitly denied in Jensen's theory of intelligence and in the theory of the "meritocracy" (Hermstein 1973; Herrnstein & Murray 1994). Thus, Hilliard (1984: 145) is correct in attributing to Jensen the error of assuming "that particular language is universal" — i.e., the thinking of IQ test enthusiasts is severely contaminated by monoglottosis.

Even the purest cases of onomatopoeia in the world's languages leave learners unable to guess either the meanings of sound sequences (or conventional gestures, as in signed languages) or the sound sequences themselves. The sounds cannot be inferred from the material states of affairs that embody the meanings (i.e., from the space-time contexts of experience) nor can the reverse be done. For instance, we cannot listen to the sound of the actual rooster out there in the barnyard and just know immediately that English-speakers will surely mimic this sound with "cock-a-doodle-do" or Spanish-speakers with "quiquiri-qui" nor when we hear someone say "bow-wow" or "arf-arf" do we unerringly infer that these are words associated with the sound made by a dog, and so on and so forth. In fact, no one can *guess with any confidence at all* just which signs out of the vast wealth of signs in any particular language will be (or are) "onomatopoetic." Neither is it possible to say for any particular language just which meanings will get onomatopoetic renderings.

In acquiring arbitrary, conventional sign-forms (where the meaning of the sign is determined by its *conventional* use), it will not help at all to listen carefully to what roosters, dogs, and bells really sound like and to imitate those sounds directly. Otherwise, the theories of language origins that John Dewey (1916) called the "bow-wow" and "ding-dong" theories would have long since won out over all other competitors not only as theories of origin but also as theories of acquisition. The trouble is that even in the extreme limiting case of the least arbitrary of all linguistic signs — namely, just those signs that do qualify as being "onomatopoetic" — the conventions of use must still be acquired in the usual ways and cannot be guessed from the surface-forms alone. It is absurd to claim, therefore, that language tasks such as those used by Binet and Simon, or in the Stanford revision, or in any other IQ test whatsoever, are *direct or pure measures of innate abilities*. They cannot be and by all evidences they are not.

Persons with monoglottosis, however, are apt to suppose that language is totally unnecessary to concepts and thought processes. This is because it really seems to them as if their own thought processes go on mostly without necessary recourse to any language or dialect — but, in fact, such processes are so dependent on proficiency in *some* language that without it they could not take place at all (also see Dewey 1916; Peirce 1898).

What's **behind** *the Binet-Simon tests?*

The idea that intelligence is hereditary did not originate with Binet, but with Sir Francis Galton (1869). Still, Binet's tests were the first attempts to measure "innate" intelligence and they preceded the first *explicit* (overtly acknowledged) attempts to measure acquired language (let alone dialect) proficiency by about half a century or more (Yerkes 1921: 355-61; Baratz 1969; Labov 1970; Carroll 1961; Lado 1957). In fact, even in the middle of the twentieth century, variance in primary language proficiency was still so completely eclipsed by the preeminent concept of "innate intelligence" as the primary source of variance in IQ tests (Jensen 1969) that the crucial role played in such tests by language would hardly be raised as a possibility. Even in the last quarter of the twentieth century when the importance of language proficiency was seriously pointed to, no less a scholar than John B. Carroll (1983) would say frankly that it had only "dimly occurred" to him that IQ tests might be measuring primary language proficiency more than anything else. Ten years later, Carroll (1993) would present a "hierarchical" view that Sternberg (1996: 11) has recently designated as "the most widely accepted view [of intelligence] at the current time." However, Carroll's hierarchical view is perfectly compatible with the language-factor theory (Oller 1996). But, alas, the time for the language-factor theory was far off in the future when Francis Galton's notions about the hereditary basis of intelligence first began to dominate the Zeitgeist of the academic world.

Based on a relatively sparse and carefully selected data sample, Galton claimed that the world's geniuses were few in number and often relatives of one another. Among the surpassing geniuses in his samples, unsurprisingly, Galton found himself and some of his own relatives, including his cousin, Charles Darwin. Therefore, he concluded, intelligence must be hereditary. With such a fulfilling concept in mind, it was easy to overlook the fact that conventional language skills are necessarily *acquired*, and that it is mainly through such *community-based conventional* sign systems that the distinct properties of human intelligence are revealed. Binet (1911: 186) himself admitted as much when he wrote:

> One of the clearest signs of awakening intelligence among young children is their understanding of spoken language. For a long time . . . our speech has affected him only by the intonation of the voice . . . The first step toward acquiring a language is its comprehension [cf. Krashen 1985; Pike 1960]. We understand the thought of others expressed in speech before we are able to express our own. Consequently, the first test is given to show that the child understands the meaning of ordinary words.

Eighty-two years later an articulate critic and reinterpreter of the concept of "intelligence," Howard Gardner (1993: xi) would still hold that IQ tests are mainly oriented toward "linguistic and logical" skills.

In addition to originating the innateness premise, Francis Galton was also the first to study twins in order to attempt an empirical test of the heritability of intelligence. He reported a number of interesting anecdotes relevant to investigations of the noteworthy shared capacities (and other personal traits) of identical twins. These would later be greatly amplified in theories of general intelligence and its inheritance (e.g. by Burt 1966, 1972; Bouchard 1981; Bouchard, Lykken, McGue, Segal, & Tellegen 1990; Jensen 1969, 1980, 1984, 1995; Herrnstein & Murray 1994). In the meantime, the specter of social Darwinism (Haller 1971) with its undisguised racist ambitions was beginning to have a worldwide influence. As Mintz (1972: 387) wrote, "*Ab initio*, Afro-Americans were viewed by these intellectuals as being in certain ways unredeemably, unchangeably, irrevocably inferior." Or, as an editor for *Science* (1972: 506) remarked, "That generation of scientists believed that no artificial process of education or forced evolution would ever enable the blacks to catch up." Ferguson (1984: 18) said, "In nineteenth-century Europe the concept of race was a preoccupation for the growing human sciences. . . . These first physical anthropologists helped to develop the concept of Aryan supremacy, which later fueled the institutional racism of Germany in the 1930s, and of South Africa today."

In fact, Galton's cousin, none other than Charles Darwin (1874) himself, was the principal source of these ideas. He prognosticated: "The civilized races of man will almost certainly exterminate and replace the savage races throughout the world. At the same time the anthropomorphous apes … will no doubt be exterminated. The break between man and his nearest allies will then be wider, for it will intervene between man in a more civilized state, as we may hope, even than the Caucasian, and some ape as low as a baboon, instead of as now between the Negro or Australian and the gorilla" (1874: 178). That scientists of the late nineteenth and early twentieth centuries generally subscribed to the racist ideas later known as "social Darwinism" is thoroughly documented by Haller (1971) in his award-winning book (also, Gould 1995).

What is more, as shown in recent critiques of the IQ enthusiasts (Allen 1995; Benson 1995; Fraser 1995; Kamin 1995b; Sedgwick 1995), though few American educators are aware of it, the main advocates of IQ testing and its recommended applications during the first four decades of the twentieth century were also associated with the eugenics movement and later with the infamous "Pioneer Fund" devoted to promoting the idea of white racial superiority (Sedgwick 1995; Kamin 1995b). In that connection, selected groups came in for criticism and were regarded as racially and genetically inferior by the elitist intellectuals led by such scholars as Brigham in the roaring '20s and destined to be approved throughout the twentieth century by Arthur Jensen, Richard Herrnstein, Charles Murray, William Shockley, and their collaborators. Until the recent brouhaha about *The Bell Curve* (at least two anthologies and hundreds of reviews have appeared) few American educators probably had any knowledge of the Eugenics Records Office (ERO) that existed in this country from 1910 until 1939, nor of its association with the Station for the Experimental Study of Evolution (SEE — which still exists) and which was founded at about the same time (Allen 1995; Kamin 1995b; Sedgwick 1995). In fact, the first director of the ERO (C. B. Davenport — see Allen 1995: 444ff) was also the first director of SEE, helping to ensure an intense relationship between the two organizations and a worldwide impact for social Darwinism. Nor are American educators generally aware that the Pioneer Fund has sponsored research from 1937 to the present day by such researchers as Jensen, Shockley, Bouchard, and others. The research sponsored by the Pioneer Fund (Sedgwick 1995), unsurprisingly, e.g., see the work of Bouchard (1981; also Bouchard *et al.* 1990), has generally supported the racial views of Darwin and Galton who claimed that blacks were closer to apes (see above).

It is easy to see Galton's influence in the use and interpretation of the IQ tests produced by Binet. Later these tests were revised, first by himself and Theodore Simon, and later by Lewis Terman and Henry Goddard at Stanford

University where they were transformed into the Stanford-Binet Intelligence Test — the cornerstone of the modern IQ testing movement and the watermark against which subsequent IQ tests would be judged (Jensen 1980: 141ff.; Herrnstein & Murray 1994: 3). In fact, the Stanford-Binet would serve as the main criterion for the creation of the U.S. Army alpha and beta tests near the close of World War I (Yerkes 1921: 563-4, 575ff.).

Binet and Simon (1905) were perfectly clear in adopting the notion of intelligence set down earlier by Galton. That is why they were emboldened to claim that their tests measured inborn "intelligence alone" (1905: 42). Yet Binet and Simon relied almost exclusively on tasks that today would be seen by any well-informed language teacher or tester as *language* tests, plain and simple. In a few cases, Binet and Simon included perceptual-motor tasks that were less obviously dependent on acquired proficiency in the particular language of the test, but as sound theory sharply demonstrates (see below), even so-called "*non*-verbal" tasks are dependent on acquired proficiency in *some* conventional language.

Among the most obvious language tests used by Binet and Simon were repetition (or elicited imitation) and ordinary dictation. Elicited imitation remains today a standard part of many so-called "verbal" IQ tests. In such tasks, a bit of discourse is presented and must be repeated verbatim. In dictation, the same sort of stimulus is presented but must be written down. Both of these tasks today have been widely recognized as tests of language proficiency for about three decades (Angelis 1972; Bachman 1990; Carroll 1961; Cohen 1980, 1994; Cziko 1983; Finocchiaro 1964; Hughes 1989; Valette 1964; Slobin & Welsh 1967; Swain, Dumas, & Naiman 1974; Natalicio & Williams 1971; Oller 1970; Savignon 1972). In fact, dictation was used as a "linguality" test to "segregate" illiterate and non-English speaking recruits from literate English speakers by the U.S. Army testers (Yerkes 1921: 348). If such tasks are presented in the primary language (i.e., the strongest and most frequently used) of the persons tested, they are rightly regarded as measures of acquired proficiency in *that particular language*. If the tasks are administered in a language *other than the primary language of the persons tested,* they are properly seen as measures of proficiency in *the foreign or nonprimary language that is used.*

For instance, suppose a dictation (or any of the other language tasks used by Binet and Simon[5]) is given to students of French in a foreign language classroom (as done by Valette 1964). It would be a measure of proficiency in French, not of native intelligence. Suppose, however, that the test were given in English, say, to foreign students at the American University of Beirut. There, the same tasks (except for the change in language) would best be regarded as tests of English proficiency, not of native intelligence. Indeed, persons familiar with the field of "language testing," which has flourished more and more since the 1960s (and has had its own professional journal, *Language Testing*, since 1984), would find it strange to call such tasks "measures of intelligence" in the sense in which Galton, Binet, Simon, Jensen, Herrnstein, and IQ testers have generally done. But imagine that a dictation (or an elicited imitation, or other language task) were given in Mandarin to persons, say, who had not had any reasonable opportunity to acquire that language. Would we regard such a task as a measure of the *native intelligence of the persons tested*? That would be absurd.

---

[5] In addition to (1) dictation and (2) elicited imitation, many other language tasks were included in the battery of tests first used by Binet and Simon. Their battery also included: (3) the repetition of numbers (a variant of elicited imitation); (4) telling what is going on in a photograph (widely used today as a measure of speaking ability in the language of the description — Engelskirchen *et al.* 1981; Palmer 1981; Choi 1994, 1995); (5) answering questions such as, "What is your last name?" or "Are you a boy or a girl?" (also commonly used by language testers in various forms as measures of language proficiency — Clark 1972: 61); (6) naming presented

Or, what if the same tasks were used with children in the process of acquiring the language of those tasks as their first or second language? Is it reasonable to suppose that a dictation (or elicited imitation, or any other language) task in Swahili, say, administered to children learning Zuni would be a valid measure of their *native* intellect? Clearly, it cannot be, because the acquisition of proficiency in any language, as we saw above, absolutely *depends* on experience

---

objects (both a method of teaching and testing in foreign language instruction — see Yerkes 1921: 357ff.; Krashen & Terrell 1983); (7) comparing two lines and saying which is longer or whether they are the same length (formidable in its language content per the instructions, and, hence, a measure of listening comprehension in addition to other sign skills); (8) comparing objects of different or the same weight (same as 7 in presenting a challenge to listening comprehension); (9) solving a puzzle; (10) counting money or determining the value of several coins; (11) executing more complex commands such as "Show me your right hand and touch your left ear" (today, language specialists would recognize this last task as falling dead center under the category that Asher (1969) termed the total physical response, a method of instilling and/or testing nonprimary language proficiency); (12) copying a phrase or sentence (involves both literacy in the language and acquired motoric skills — one might try doing it, for instance, in a language not yet known to the test-taker, say, Arabic, or Chinese); (13) reading aloud and recalling points of information (ubiquitously depends on acquired language proficiency); (14) giving the day of the week and the date — day, month, year (again, one might try it in a foreign language); (15) making change with money; (16) giving abstract definitions of words (essentially involves acquired conventional uses of the language in question); (17) arranging different weights in a series from least to greatest (listening comprehension as well as acquired perceptual and motor skills are involved); naming months of the year (similar, to 14 above); naming pieces of money (similar to 6 above); (18) using three words to construct a sensible noncontradictory sentence or placing disarranged words in order (used in many variations to instill or test language proficiency — De Berkeley-Wykes 1983); (19) answering questions about physical or social problems — e.g., what to do when you miss a train, or why a person should be judged more by actions than by words (used under the guise of psycho- and socio-drama both as a method of teaching and testing language proficiency-cf. Scarcella 1978; Stern 1980; Di Pietro 1981, 1982); (20) saying what is wrong with a sentence such as "You are to be hanged at dawn: Let this be a warning to you!" (involves inferences about the presuppositions and implications of conventional uses of linguistic signs); (21) producing 60 different words in three minutes (involves a peculiar and unnatural application of language skills and almost nothing else); (22) giving abstract definitions of concepts such as "charity," "justice," and "goodness" (try it in a foreign language); (23) saying how a cut paper will look when it is unfolded (a task formidable mainly in its requirements on linguistic production); (24) rotating a right triangle to coincide with another by following instructions (listening comprehension); and (25) telling what the difference is between paired attributes such as "pleasure/happiness" (essentially involves acquired language proficiency and is similar to 22).

with the target language in some of its contexts of use. It is only a little less absurd, even when the tasks are in the child"s primary language (or preferred dialect) to suppose that they are direct, valid, and pure measures of *native* intelligence. But this is exactly what the vocal IQ proponents have claimed by their innateness premise for essentially *all* so-called "verbal" IQ tests (voices of caution such as Carroll 1993; Gardner 1983, 1993, 1995; Linn 1989; Mercer 1973, 1984; Sternberg 1985, 1996 — to mention only a few — notwithstanding).

Examining the army alpha tests (precursors to verbal IQ tests)

The U.S. Army alpha tests were steeped in the legacy of social Darwinism and were destined to become the precursors of "group verbal" intelligence tests in general. The Army tests were highly praised by Brigham (1923: xx): "The army mental tests give us an opportunity for a national inventory of our own mental capacity, and the mental capacity of those we have invited[6] to live with us.... We find reported... the intelligence scores of about 81,000 native born Americans, 12,000 foreign born individuals, and 23,000 Negroes. From the standpoint of the numbers examined, we have here an investigation which, of course, surpasses in reliability all preceding investigations, assembled and correlated, a hundred fold. These army data constitute the first really significant contribution to the study of race differences in mental traits."

Linking his argument with race was the whole purpose, evidently, of his investigations. He concludes in the end, "According to all evidence available, then, American intelligence is declining, and will proceed with an accelerating rate as the racial admixture becomes more and more extensive. The decline of American intelligence will be more rapid than the decline of the intelligence of European national groups, owing to the presence here of the Negro" (ibid.: 210). He laments that there were already "10,000,000 Negroes" in the country and that they had been mixing with whites with "alarming rapidity since 1850" (ibid.: 210).

Bearing in mind the pungent odor of the entire eugenics movement which spurred the Army testers, and recalling its roots in the thinking of Galton and Darwin, consider what the eight "verbal" or "alpha" tests, as they were called, consisted of:

1. The first, "Oral Directions," was "a series of commands or directions which were to be executed quickly" (Brigham 1923: 3) and each of these 12 items was accompanied by a numbered drawing consisting of either geometrical figures, letters, numbers, or some combination of these. For instance, item 4 presents a figure consisting of three overlapping figures, a circle, square, and triangle and

---

[6] Evidently the taking of black slaves from Africa, halted in this country by the bloody Civil War, was regarded euphemistically by Brigham, a declared eugenicist, as a mere "invitation" for black slaves to join us over here.

asks, "Make a figure 2 in the space which is in the circle but not in the triangle, and also make a cross in the space which is in the triangle and in the square" (ibid.: 3). To see that this is primarily a language task along the lines of Asher (1969), it is helpful to imagine trying to do it when the instructions are given in, say, Hottentot, Igbo, or American Sign Language. Also it is pretty ridiculous to say that the main difficulty of the task is in "abstract thinking" (per Jensen 1980: 132) rather than in "the use of language." The two cannot reasonably be separated in such tasks. It is curious that 73 of the individuals in the experimental sample of 1,047 white recruits, all supposedly native speakers of English, could not answer a single question correctly on this test — i.e., they all obtained scores of zero.

2. The second test consisted of what were called "Arithmetical Problems" and all 20 of these were formulated as word problems, e.g., item 4 says, "Mike had 11 cigars. He bought 2 more and then smoked 7. How many cigars did he have left?" (Brigham 1923: 9). While there is no doubt some arithmetic here, and though some of the items could be most easily solved by translating them into simple linear equations, the principal difficulty is in understanding the propositional relations expressed in the language of the test. Here 66 out of the 1,047 subjects failed to get a single correct answer.

3. The third Army alpha test contained 16 items supposedly aimed at "Practical Judgment," each one in a forced-choice format with three choices for each question: Item 15 says, "If you are held up and robbed in a strange city, you should
- apply to the police for help
- ask the first man you meet for money to get home
- borrow some money at a bank" (ibid.: 13).

The task could be described as determining the most plausible written sequitur for a given proposition or state of affairs described in writing. More than 15 percent of the sample (163 subjects) missed all these items.

4. The fourth test, containing 40 items, was called a "Synonym-Antonym" task and required subjects to say whether a given pair of words was the same or opposite, e.g., in item 1, "yes" and "no" are given and in item 26 "vestige" and trace" (ibid.: 18), and the subject must say for each pair whether they are the same in meaning or opposite. In this case, the score assigned was the number of correct answers minus the number of incorrect answers (ibid.: 16). This task is the precursor of many similar vocabulary tests which are included not only in nearly all so-called "verbal" IQ tests today, but also in most reading tests and in many general achievement tests and in most language proficiency batteries (e.g. the Test of English as a Foreign Language). About 37 percent of the experimental subjects failed this test entirely.

5. "Disarranged Sentences" (24 items), the fifth test in the alpha series, consisted of a series of items where the subject must unscramble a given word-sequence to make sense of it and then say whether it is true or false, e.g., item 19 says, "men misfortune have good never" which is to be unscrambled to get "Good men never have misfortune" which is to be judged false. Unscrambling tasks are commonly used in second language tests with words, phrases, or sentences (cf. De Berkeley-Wykes 1983), and so are true-false judgments (e.g. Clark 1972, 1979; Clark & Grognet 1985; Heaton 1975). This test did not differentiate well at the bottom of the distribution because nearly everyone in the bottom quarter of the distribution (244 individuals; Yerkes 1921: 577) obtained a score of zero. The idea of unscrambling words in a list such as "begin a and apple acorn ant words with the" to get "The words apple, acorn, and ant begin with 'a'" does not much resemble ordinary language use.

6. The sixth alpha task consisted of 20 items called "Number Series Completion." In each of these a progression of six numbers is given and followed by two blanks to be filled in. For instance, the first item in the series can be answered by merely counting by ones: 3 4 5 6 7 8 _ _ (ibid.: 24), i.e., the blanks should be filled by 9 and 10, but the twentieth in the series requires a progression formed by first multiplying by 2 and then adding 2: 3 6 8 16 18 36 _ _, where the blanks should be filled by the numbers 38 and 76. While this task is one that Brigham may have thought involved no English at all, the instructions for it are formidable and a little stilted: "In the lines below, each number is gotten in a certain way from the numbers coming before it. Study out what this way is in each line and then write in the space left for it the number that should come next. The first two lines are already filled in as they should be" (ibid.: 22). It is interesting that Brigham feels compelled to note that "in the final alpha revision, four samples were included, and the instructions were simplified verbally and read very slowly" (ibid.). He tries to explain the fact that nearly 16 percent of those tested (177 out of 1,047) could not understand the instructions by saying, "Although there were many zero scores in our experimental group . . . , there were probably no more zero scores than might have been expected when we consider that the mere understanding of what was wanted required considerable intelligence" (ibid.).

7. The seventh test was called "Analogies" and consisted of forced-choice items. For instance, in item 16, "egg—bird:: seed— " is followed immediately by four choices in bold print, in this case, "grow plant crack germinate" and subjects must choose the one that best completes the analogy ("plant" in this case). In this particular test, although the task is clearly one that focuses most of the energy on mental operations that are utterly dependent on the language of the task, the instructions are substantially more formidable than any one of the items: "In each of the lines below, the first two words are related to each other

in some way. What you are to do in each line is to see what the relation is between the first two words, and underline the word in heavy type that is related in the same way to the third word. Begin with No. 1 and mark as many sets as you can before time is called" (ibid.: 25). The difficulty of the instructions is shown in the fact that 284 of the subjects tested failed to answer even a single item correctly (27 percent of the sample). Brigham was puzzled: "The analogies test is the most effective in the entire series in differentiating officers from men. For some reason, not understood, it does not rank high in differentiating feebleminded from enlisted men" (ibid.: 27). But, of course, those who failed to get a single correct answer had to be at the bottom of the distribution and to end up among the "feebleminded" according to Brigham. However, if becoming literate can cure "feeblemindedness," his worry about declining intelligence was wasted. What was needed was merely to improve English language proficiency, including reading and writing skills (for evidence that this can be done and that it shows up in "IQ" scores, see Rainey 1992).

8. The last test in the alpha series was titled "Information" and consisted of items in a forced-choice format such as item 26, "The author of "Huckleberry Finn" is followed by the choices in bold face "**Poe Mark Twain Stevenson Hawthorne**" and item 40, "An irregular four-sided figure is called a **scolium triangle trapezium pentagon**" (ibid.: 29). In defense of such items being included among the alpha tests, Brigham said, "Approximately one-third of the times [sic, should be "items"] test for vocabulary rather than information in the literal sense. If a person, for instance, knows what a Zulu, or a Korean, or a Hottentot, or a Kaffir, or a Papuan is, he very obviously knows the number of his legs" (ibid.: 30). But suppose the person did not know the vocabulary item or bit of information coded in English, or Zulu, or Korean, or Hottentot? Would that mean such a person also did not know the number of his or her legs? The fact is that what Brigham evidently regards as razor-sharp logic — the sort that would guide the intelligence test enthusiasts until the present day — was grounded in extreme monoglottosis. He was, as Eysenck, Hermstein, Jensen, Lynn, and others would be, quite unaware of the crucial role played by language proficiency in accessing information. Out of the 40 items included in the "Information" test, on the average the experimental sample answered only 15 items correctly (out of 40) and 159 subjects (Yerkes 1921: 578) scored either a zero or one (a point was subtracted for wrong answers). Brigham supposes all these individuals were "feebleminded" (Brigham 1923: 31).

What the foregoing examination shows is that the U.S. Army alpha tests, which would later serve as a touchstone for all the group tests to follow (e.g., the Otis tests, Lorge-Thorndike, Wechsler, Kaufman, etc.), were as their successors are, mainly tests of proficiency in the language of the tests. Of course, they could not be otherwise and still be called "verbal" tests. The travesty is that instead of supposing that they are measuring proficiency in English (however well or badly), it was supposed that they measured innate intelligence — and there can be no doubt that this interpretation was rooted in the eugenics objective of racial "purification."

No one would have become alarmed if it had been pointed out that about 30 percent of U.S. Army recruits were barely literate, or that the average reading ability of all those tested was equivalent to that of, say, an eighth grader, but many did become alarmed when some concluded (Stoddard 1922) that the average innate intelligence of American Army recruits was equivalent to that of a 14-year-old child. In response, Walter Lippmann (1922) quipped that "Mr. Stoddard's remark is precisely as silly as if he had written that the average mile was three-quarters of a mile long" (Jacoby and Glauberman 1995: 561).

The touted "nonverbal" IQ tests

The common defense against the foregoing arguments has been to take refuge in so-called "nonverbal" IQ tests. The crux of the case, therefore, comes down to them. The question is *are "nonverbal" IQ tests really "language free" and "dialect-free" measures of "innate intelligence"? Do they justify the theory of the "meritocracy" and its unsavory racial conclusions?* There is no better illustration of the effect of monoglottosis among the vanguard of IQ enthusiasts than can be seen in the instructions of the U.S. Army beta "intelligence" battery and the remarkable and convoluted attempts to deny that the beta tasks demanded proficiency in the English language. Some IQ enthusiasts will try to defend against these criticisms by saying that more recent nonverbal tests have escaped the difficulties of the beta tests, but we will see below that this is not true and indeed we will demonstrate by strict theoretical reasoning that it can never be true in principle.

We will see below that pantomimed instructions can *never* remove the need for access to abstract ideas through *some particular language*. The beta series is therefore not exceptional and it is historically important because it provided the foundation for all of the subsequent so-called "nonverbal group" IQ tests, e.g. Raven's Progressive Matrices developed in the 1930s, Cattell's Culture-Fair Tests developed at about the same time, and so forth. The beta tests were called "nonverbal" because they supposedly eschewed all dependence on any language and they were called "group" tests because they were designed to be administered to 25 to 100 persons at a time (Yerkes 1921: 368). Since then, Jensen, Herrnstein, and others have persisted in claiming that Raven's and Cattell's tests are quite pure measures of *g* — i.e., of innately determined general intelligence. The plethora of derived "nonverbal" tests have variously been called "performance" (as in the Wechsler scales for adults and children, the Kaufman tests, etc.), "culture-free," "culture-fair," "unbiased," "culture-reduced," or "language-free" tests. Of such "nonverbal" tests, Jensen (1980) would have to admit that "the directions are usually *verbal* [italics added]" by which he meant *in English,* "but," he would protest immediately, "a good nonverbal test begins with items of any particular type that are so simple that virtually all subjects can catch on to the requirements of the task without verbal instructions, or with pantomimed instructions by the tester" (ibid.: 132).

This is merely a paraphrase of the claim originally made on behalf of the U.S. Army beta tests devised to test "illiterate and non-English speaking" recruits near the end of World War I. Literate English-speakers were tested on the alpha tests (commonly referred to as tests of "literacy" and "linguality" in English, Yerkes 1921: 327ff.), but it was found that "supplementary tests" were needed "to give those handicapped by language difficulties [i.e., persons speaking *any language other than English*] a real opportunity to show their ability" (Yoakum & Yerkes 1920: 10). The whole idea of the beta tests, therefore, and all of their successors was that they were supposed to be given without recourse to any particular language. Carl C. Brigham,[7] one of the psychologists who played a significant role in the development, use, and interpretation of the U.S. Army IQ tests, described the beta tests as "seven different sorts of tests, none of which involved the ability *either to read English or to understand spoken English* [italics added], the tests consisting of pictures, designs, etc., and being given by instructions in pantomime" (Brigham 1923: xxii).

Therefore, the crucial question for the beta tests, and their successors, is whether the pantomime approach really works or not. Yoakum and Yerkes (1920), in the authorized government manual for administering the alpha and beta tests, warned test administrators that "the subjects who take this examination [beta] sometimes sulk and refuse to work . . . . " (ibid.: 80). Yerkes (1921: 379) said that "the main burden of the early reports was . . . that the most difficult task was 'getting the idea across.'" He admitted that "a high percentage of zero scores . . . was considered an indication of failure to 'get that test across'" (ibid.). He quoted one of the users as saying, "There is no doubt that beta is many times harder to give than alpha and requires constant effort from everyone concerned. Nothing can be more fatal to it than the alpha method of giving" (ibid.: 385). Yoakum and Yerkes (1920: 80) said that "with the exception of the brief introductory statements and a few orders, instructions are to be given throughout by means of gestures instead of words. These gestures accompany the samples and demonstrations and should be animated and emphatic." It was also admitted that "variations of procedure are more likely to occur in beta than in alpha . . . E. [the examiner] should especially guard against using more or fewer gestures or words for one group than for another. Oral language should be rigidly limited to the words and phrases given in the procedure for the different tests . . . . Both examiner and demonstrator must be adept in the use of gesture language .... One camp has had great success with a 'window seller' as demonstrator. Actors should also be considered for the work" (ibid.: 81).

---

[7] Brigham would later become the first executive secretary for the College Entrance Examination Board at Princetown, the designer of the Scholastic Aptitude Tests, and for a time the elected secretary of the American Psychological Association (Kamin 1995b: 495-6).

So that readers may judge for themselves just how much language was required in the beta tests, here are the general instructions (abbreviated some) together with those that were common to each of the seven beta tasks from the authorized manual. I suggest substituting Vietnamese, Mandarin, Mongolian, or any language unknown to the reader for the remarks in English (the portions within quotes) and trying to imagine whether or not the instructions, gestures, and pantomime would be comprehensible:

> E. [the examiner] should say "Here are some papers. You must not open them or turn them over until you are told to." Holding up the beta blank, E. continues: "In the place where it says name, write your name; print it if you can. (Pause.) Fill out the rest of the blank about your age, schooling, etc., as well as you can. If you have any trouble, we will help you." ... After the initial information has been obtained, E. makes the following introductory remarks: "Attention! Watch this man (pointing to demonstrator). He (pointing to demonstrator again) is going to do here (tapping blackboard with pointer), what you (pointing to different members of group) are to do on your papers (here E. points to several papers that lie before men in the group, picks up one, holds it next to the blackboard, returns the paper, points to demonstrator and the blackboard in succession, then to the men and their papers). Ask no questions. Wait till I say 'Go ahead!'"... [For each of the seven tests that follow these general instructions, the specific instructions include:] "Now turn your papers over. This is Test X [a particular number is given] here (pointing to the page of record blank). Look. Don't make any marks till I say 'Go ahead.' Now watch .... No—no .... Good .... Look here. All right. Go ahead. Do it (pointing to the men and then to books). Hurry up. Do it, do it, hurry up, quick .... Stop! Turn over the page to Test Y [the next in the series]" (Yoakum and Yerkes 1920: 81-2).[8]

Difficulties were evident in the high percentage of zero scores attained by the specially selected "experimental group" of 1,047 recruits (see Brigham 1923: 3-58; Yerkes 1921: 563-4 and 573-657). This group supposedly contained only English-speakers. This was regrettable in view of the fact that the beta tests were specifically intended for use with persons from non-English speaking backgrounds, or for illiterates of any background. Therefore, the tests should have been tried out on persons who were illiterate and/or who did not understand English. In spite of this, the percentage of subjects who received a zero score on one or more beta tasks ranged from 2 percent (19) to 10 percent (105) compared with 7 percent (73) to 37 percent (393) who scored zeroes on alpha tests (Yerkes 1921: 577-81).

---

[8] Also, it should be noted that the instructions to the beta tests would make a very hard first lesson for any foreign language. The notion that students could just walk into the classroom and immediately understand such instructions, even when accompanied by "emphatic" gestures and "pantomime" will be implausible to any practicing foreign language teacher. The beta instructions require some giant steps of nonprimary language acquisition.

If for no other reason, it should have been obvious to the test designers from correlations alone that the language of the instructions and in the items themselves may well have been the most important factor in the alpha and beta tests. The correlation of the raw scores on the alpha battery with the beta battery for the "experimental group" (*n* = 1,047) was reported at .811 ± .009 (Yerkes 1921: 392) and with the raw scores on the Stanford-Binet for the sample of 653 experimental cases (perhaps with the non-English speaking and illiterate persons purged) it was .727 ± .012 (ibid.: 390). Interestingly, the alpha battery correlated at about the same level with the "Devens literacy" test, .831 (*n* = 289 whites at Fort Meade), .813 (*n* = 375 whites et Camp Meigs), .850 (*n* = 380 whites at Camp Lee) as it did with the beta tests .793, (ibid.: 373), and somewhat better than it did with the Stanford-Binet (where *r* = .727).[9] Also, the separate "linguality" tests (individual language tests modeled after the Stanford-Binet) and group tests (modeled after the beta tests) correlated with each other at higher levels, .793 for 630 mixed cases of both English and non-English speakers, than the Stanford-Binet correlated with the beta tests (.727).

Besides the cross-overs just noted, where so-called IQ tests correlate more strongly with "language proficiency" measures (and vice versa) than with each other, there are other more recent studies that could be cited. To mention just one, Kelly (1965) found a correlation of .85 between the Brown-Carlsen Listening Comprehension test and the Otis Test of Mental Abilities while the Brown-Carlsen correlated at .82 with the Sequential Test of Educational progress, another test aimed at listening comprehension, which nevertheless correlated at .85 with the Otis IQ test. Carroll (1972) commented that it was "rather disturbing" that "the various tests of 'listening ability'" tend to show no higher correlations among themselves than they show with reading and intelligence tests" (Carroll 1972: 2; also see Todd & Levine 1994: 105). *Oddly, however, the tendency has been to question all other tests rather than to doubt the innateness premise associated with the IQ tests.* However, as we have already seen, the innateness premise cannot be correct.

We have already examined the tasks used by Binet and Simon as well as those of the U.S. Army alpha (verbal) and beta (nonverbal) IQ tests and we have seen that those tests are clearly dependent on acquired language proficiency. I argue that the *language factor*, therefore, is the main source of variance (i.e. the statistically observed differences in scores) in IQ tests[10] and is owed to: (1) differences in the languages and dialects used by testers in instructions and items; (2) differences in the accessibility of those languages and dialects to the persons tested; and (3) consequent differences in the accessibility of the information to be manipulated or supplied in the tests. The language factor (as just defined above) not only constitutes the main source of variance in "verbal" IQ scores (as has already been shown above), but more importantly, is also crucial to "*non*verbal" IQ scores. In short, the language factor *is the primary basis for Spearman's* g.

---

[9] This may well have been owed in part to the time-lapse between the group testing for alpha and the individual testing required for the Stanford-Binet.

[10] The argument also extends to achievement tests, pupil rating inventories, and even to most personality inventories as first suggested nearly twenty years ago (Oller & Perkins 1978).

As noted from the beginning, in the final analysis my entire argument hinges upon the validity or lack thereof of claims made for "nonverbal" IQ tests. In the following section, we examine those claims closely and demonstrate by strict logic that "nonverbal IQ" tests cannot, in principle, ever escape dependence on the primary language abilities of the persons tested. The argument is abstract and the task is not made easier by the accumulated debris of almost a century of misinterpretations of IQ test scores, but the most crucial aspect of the *language-factor theory* is what follows. It aims to be strictly logical and I have endeavored to present it in easy steps so that readers may judge the case for themselves.[11]

The Basis of "*Non*verbal" IQ Tests

It is commonly *claimed* by the staunchest defenders of the innateness premise that the strongest predictors of *non*verbal IQ scores are *verbal* IQ scores. This is not always correct, but it would very nearly be correct if "proficiency in the primary language" were substituted for "verbal IQ." For instance, Jensen (1980) writes: "In a large sample of school children who had taken the Lorge-Thorndike Intelligence Test, for example, we found that the correlation of a reading test (Paragraph Meaning subtest of the Stanford Achievement Test) with verbal IQ is .52; but the reading scores correlate almost as highly (.47) with the nonverbal IQ, which requires no reading at all. The correlation between the verbal and nonverbal IQs is .70 in this sample" (Jensen 1980: 132). From all of this he concludes, "obviously the verbal IQ reflects reading ability per se to only a relatively small degree" (ibid.). The implication that he draws is that verbal IQ *and* nonverbal IQ scores both mainly show innate ability. He goes on to point out in a parenthesis that "the partial correlation between reading and verbal IQ, holding nonverbal IQ constant, is .29" (ibid.). The implication here is that verbal IQ scores only partly reflect reading proficiency and that they must measure something more, namely, according to Jensen, innate ability.

---

[11] Perhaps it should be noted here that in spite of the strong claims made for the logical underpinnings of the language-factor theory, and in spite of the interpretations of empirical evidences offered here, I should not wish anyone to think I am saying there is no room for disagreement. On the contrary, my object is to present the theory and some of its empirical results as clearly as possible for the purposes of careful examination, further testing, and the most intense scrutiny possible from all conceivable angles. The vast literature on IQ testing is merely one among many sources of empirical data against which the underpinnings of the theory can be examined. Indeed, owing to its strict logical basis, the most obvious test is to examine the theory for any logical inconsistency. Just one of those could be fatal to the entire enterprise. Therefore, the strength of the claims made for the theory are an invitation for sharp and clear disproof, whether empirical or logical, which is the greatest service that can be done for any theory. Of course, readers are also welcome to disagree merely for the sake of discussion, so that together we may endeavor to clarify whatever element of the argument may require it.

Mercer (1984, citing Wechsler 1974), though aiming mainly to criticize the standard interpretations of IQ scores, also shows substantial correlations between the Wechsler Intelligence Scale for Children-Revised (WISC-R) verbal and performance (nonverbal) scales for white elementary students, $r = .57$ ($n = 668$); for Hispanic elementary school students, $r = .59$ ($n = 613$); and for a similar group of blacks, $r = .61$ ($n = 619$). She also gives a factor analysis for the same correlation data by group showing similar patterns of loadings on a first principal component for each of the groups and on two orthogonal factors for each of the three groups. Interestingly, the verbal factor showed significant loadings from the nonverbal (performance tests) and the reverse was also true (Mercer 1984: 305-7). Therefore, the defenders of IQ tests contend, *both* kinds of IQ tests *must be* measuring genetically determined abilities, but as Mercer (1984) says, "In every case, the verbal factor [i.e., proficiency in white English] accounts for an overwhelming percentage of the variance" (ibid.: 308). In fact, the first principal component accounted for 85 percent of the variance of her white subjects, 87 percent for Hispanics, and 86 percent for blacks. For all three groups, the highest loadings on the first principal component was "Vocabulary" subtest of the WISC-R. Clearly, the language factor is the principal element.

The alternative view, i.e., the innateness premise advocated by Jensen and others, has nothing to support it and it leaves the Kelly (1965) data discussed above and the other cross-overs discussed there unexplained. Moreover, the innateness premise conflicts with the much simpler language-factor theory that accounts for all of the data already presented as well as the "unsolved mysteries" to be considered below.

In fact, the usually significant and sometimes substantial correlations commonly observed between language proficiency tests and *non*verbal IQ tests (roughly from about .40 to .85 — Jensen 1980; Mercer 1984; Oller 1983; Oller & Perkins 1978, 1980; Stump 1978; Yerkes 1921) ought, at the very least, to suggest the hypothetical possibility that *the so-called* "non*verbal" IQ tests also may depend on skills developed in acquiring a language.* Moreover, correlations of "*non*verbal IQ" tests are consistently stronger with tests in the primary language of the test-takers than with so-called "*verbal* IQ" tests given in absolutely *any* language *other than* the primary language of the examinees. This fact ought, at the very least, to suggest the hypothesis that *the main factor measured by the* "non*verbal" IQ tests may really be the primary language proficiency of the test-takers.* Furthermore, both of these hypotheses which are suggested by the empirical research on IQ tests can be deduced from the language-factor theory. I already showed above why all "verbal IQ" tests depend on access to the conventional signs of the language used in those tests, and it remains now to show that all "*non*verbal IQ" tests depend likewise on access to abstract ideas attainable only through the conventional signs of *some* particular language.

The strict theoretical argument demonstrates that access to any conceivable abstract idea (i.e., any comprehensible or translatable thought about objects, relations between objects, relations between relations, etc.) absolutely requires conventional signs of the linguistic kind.[12] Perceptual icons (i.e., sensations derived from seeing, hearing, touching, tasting, or smelling material objects or states of affairs) are inadequate because they are essentially private and non-transferable from one person to another. Similarly, bodily movements (i.e., deliberate motoric signs such as pointing) cannot assure adequate communication between persons because of the universal degeneracy of indices. It can be shown that all bodily movements are a proper subclass of indices and that the entire class is degenerate in a special way. Pointing to a square on a sheet of paper, or a blackboard, or an orderly, for example, or to any conceivable object that might be pointed to leaves the observer well short of knowing for sure what is being pointed at. Anything along the line of pointing is a possible candidate. Is it the square? The paper? The desk where the paper is located? Even touching the object directly does not identify it unambiguously. Is it the square itself that is being touched or the line that defines the square or a particular corner of the square, or the point where the lines join at that comer, etc., etc., *ad infinitum*?

---

[12] The theoretical argument consists ultimately of a series of tightly linked logico-mathematical proofs. The latter, if executed properly, depend on nothing but the sort of consistency sought by mathematics in general. Such consistency cannot be rejected without both introducing an inconsistency and demolishing the whole house of mathematical reasoning. Omitting details, let me sum up the argument and show its application to the special problem posed by "nonverbal" IQ tests. First, we know that the existence of meaningful signs cannot reasonably be doubted or denied because every argument suited to express such a doubt or denial is obliged to overtly employ the sorts of signs it claims to doubt the existence of. Neither can the existence of material objects be doubted or denied because every conventional sign is manifested only in one or more discrete material objects (e.g., such as are constituted by the letters, words, punctuation marks, etc., of this sentence). Nor can the existence of conventional relations between signs and material objects be reasonably doubted or denied. If someone should attempt to frame such a doubt or denial, the denial itself would turn out to involve conventional signs linked to that individual in a conventional way or else the argument would be incomprehensible. By these arguments, the foundation for signs that are truly related to the material objects of someone's experience is laid. Should any skeptic wish to argue against the actual existence any such true narrative signs (i.e., signs truly connected to genuine experience), that skeptic would be obliged to produce one or more exemplars. This is because the skeptic's argument at a minimum would have to take the form, "I demonstrate by these signs, etc., etc., that true narrative signs either do not exist at all or that their existence is doubtful." But, every such argument as formulated is a true narrative sign because by its very form it is truly related to the experience of the person uttering it, not to mention that of anyone who may comprehend it. Therefore, true narrative representations cannot be reasonably doubted or denied existence.

From such a foundation, it is possible to show by a strict progression of arguments that the meaning of all possible signs is owed to signs of the true narrative kind. Thus, any general theory of sign acquisition, or intelligence, need only attend to signs of the true narrative kind because every other kind is parasitic and derivative. Therefore, a strict theory of abstraction can be developed showing a certain progression in the natural development of true narrative signs in the experience of any child. In fact, a strict hierarchy of true narrative signs follows. The most basic of these are perceptual signs arising from material forces (from light, heat, mass, etc.) emanating from material objects linked ultimately with the child's own body through indices consisting of bodily movements. However, it can be rigorously demonstrated that the abstract meanings of nonverbal IQ test items, as agreed upon by testers and persons answering those items correctly, can only be attained (and agreed to) through the shared conventional signs of some language. In fact, abstract meanings only become accessible to any developing sign-user through the conventional linguistic acts of one or more sign-users who are part of the sign-community. Access cannot be attained through perceptual signs exclusively because these signs do not have a shared conventional element. Access cannot be attained through non-conventionalized gestures or pantomime either, because these signs also lack the necessary conventional element. That is, nonconventionalized gestures and pantomime cannot express abstract ideas unambiguously. Thus, the case is made.

The reason for the degeneracy of every index stems from its resemblance to a line. Although any movement along a line segment logically has only two ends, owing to the natural extendedness of the line saying just where either end really is pointing is problematic. With respect to the *origin* of the gestural index (where it begins), so long as only one person is doing the gesturing (or "pantomime"), there is not much problem, but when it comes to the *terminus* (where it ends), the problem is acute. Furthermore, in giving gestural and pantomime instructions for completing nonverbal IQ tests, the experts have sometimes recommended having more than one person do the pantomime. In these cases, the origin of relevant gestures is also rendered indeterminate. The subject may be looking at the wrong person and there is no way from gestures alone to determine which person is supposed to have the floor. Is the test administrator, for instance, pointing to the orderly or the blackboard? Or is the orderly directing the test-taker's attention somewhere else?

The IQ testers have generally supposed that unconventionalized gestures or mere "pantomime" could be added to perceptual signs in such a way as to obviate the need for any linguistic signs whatsoever. But this is not possible. Only conventionalized symbols, concepts of the kind that require *conventional linguistic signs to define them,* can escape the peculiar degeneracies of icons and indices just pointed out. Moreover, it is simply an error grounded in monoglottosis to claim as Jensen (1980: 132) does that "virtually all subjects can catch on to the requirements of the task without verbal instructions." This is not true for the simplest conceivable designation of any object and it is certainly not true for propositionally complex reasoning of the sort that enters into such tests as Raven's Progressive Matrices or the Cattell Culture-Fair Tests, or any others. The latter nonverbal tests are commonly cited by Jensen and others as paradigm examples of the very best "nonverbal" tests ever produced (Jensen 1980: 648ff.).

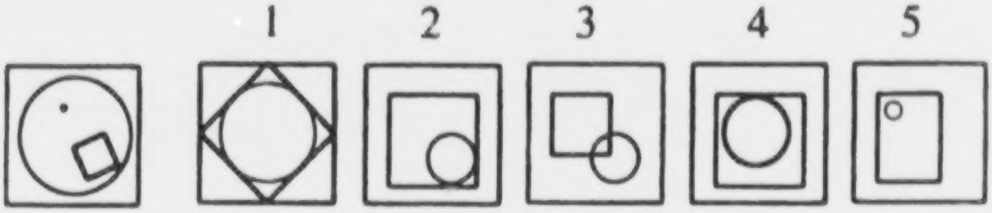
To demonstrate practically that the instructions cannot be pantomimed adequately by any stretch of the imagination, consider the "sample item" in Figure 1 which is taken from the published introduction to the Cattell Culture-Fair Intelligence Test. As an imaginary experiment, the reader may want to try to guess what the instructions for completing the item may be. A more difficult challenge would be to invent a pantomime presentation to show a group of

---

[13] The actual instructions are to find the place on the right where the dot can be placed in the same relation to the circle and square that it has in the figure on the left.

uninformed subjects how to complete the item.[13] Finally, as a clincher, have someone translate the instructions into Vietnamese, Mandarin, or Navajo, and see if anyone who does not know the language of the instructions can figure them out from gesture and pantomime.

Figure 1: A sample item from the Cattell Culture-Fair Intelligence Test



Even in open-ended IQ tasks (as seen in certain individually administered tests, e.g., the Goodenough-Harris Draw-a-Man Test) the suitably intelligent sign-user is expected to be able to produce some response at a degree of socially recognizable appropriateness greater than zero. Any such claim in behalf of the validity of any such IQ test, however, is tantamount to the overt confession that every IQ item of that kind *must* rely principally on acquired proficiency in some language in order for the persons tested to gain access to the meanings at stake. As a consequence, the solving of any meaningful (potentially valid) nonverbal IQ item depends on access to signs that are both general *and* conventional. But signs with these properties, owing to the arbitrariness of their surface-forms (which is strictly demonstrable for all meaningful conventional signs), are indistinguishable from the semantic content of linguistic signs and are accessible only through linguistic signs.

Some "Unsolved Mysteries" of IQ Testing

There are a number of "mysteries" in the IQ literature that are often singled out by supporters or detractors of the innateness premise and its meritocracy theory. Here are nine of them: (1) Why are minority language children so commonly over-represented in classes for the mentally retarded, learning disabled, and the like (Jitendra & Rohena-Diaz 1996)? (2) Why are minority language children so seriously under-represented in classes for the gifted (Gersten & Woodward 1994; Ortiz 1987)? (3) Why are there nearly universal discrepancies between so-called "verbal" IQ (VIQ) and so-called "nonverbal" or "performance" IQ (PIQ) scores for minority language children — i.e., why if both of these types of tests are measures of the same innate intelligence (tapping into the same *g*) do minority language groups typically show higher PIQ scores than VIQ scores when the VIQ scores are based on English-language tests (Figueroa 1989; Valdes & Figueroa 1994)? And (4) how are the exceptional cases to be explained? Why is it that some Asian[14] children seem to acquire English so rapidly that the typical VIQ/PIQ discrepancy disappears early in their experience in the U.S. (Gordon 1980; Jensen 1984; Figueroa 1989)? Similarly, (5) why is it that Afro-Americans in the literature sometimes do about equally well on VIQ and PIQ tests and still appear to be about 15 points below the mean for supposedly comparable white groups (Jensen 1974, 1980, 1984; Mercer 1973, 1984)? (6) Why does this same sort of difference also appear for Irish subjects as contrasted with English subjects (Lynn 1978, 1979)? (7) Why do deaf children of deaf parents score higher on IQ tests than deaf children of hearing parents and comparable groups of hearing children? In fact, the observed differences average about 13 points in favor of deaf children of deaf parents (Gallaudet Research Institute 1987; Isham & Kamin 1993). But why is this? In this same connection, (8) what is to be made on the long-standing controversy over "Spearman's hypothesis" (Spearman 1927: 379) that *g*-loadings for tests applied to minority groups will be correlated with the majority minus minority mean scores on those same tests (Braden 1989; Isham & Kamin 1993; Jensen 1980)? Why should this be so? And, why does it hold or not hold for certain samples of deaf children depending on the criteria for sample selection — e.g., deaf children of deaf parents (where it does not hold) versus deaf children of hearing parents (where it does seem to hold; see Isham & Kamin 1993)? Finally, (9) how can we explain the reported declining IQ of immigrant populations that supposedly occurred in the U.S. in the first quarter of the twentieth century (Brigham 1923; as supposedly documented by the Army alpha and beta IQ tests)?

I contend that the language-factor theory affords a straightforward solution to each of these mysteries and that the false innateness premise of the IQ testers is the bogus source for each of them. Furthermore, as we examined the history of IQ testing above, we found that Galton, Darwin, Brigham, Eysenck, Gordon, Herrnstein, Jensen, Lynn, Murray, and those of whom these have approved, had a racial axe to grind from the start. It is that underlying commitment to a racist outlook that accounts for their stilted, ad hoc, and implausible explanations of the foregoing mysteries by contrast with the simpler language-factor theory. I agree in the final analysis with Peirce (1908) who sided with Galileo concerning *il lume naturale* (quoted in Hartshorne and Weiss 1935: 325). According to Galileo, "of two hypotheses, the *simpler* is to be preferred" and Peirce came to understand this to mean "in the sense of the more facile and natural" (ibid.: 326). For the sake of convenience, I will show how each of the alleged "unsolved" mysteries can each be "solved" by the language-factor theory. In fact, the "mysteries" are not so much solved as shown to be artifacts of the innateness premise and the meritocracy theory which are both false.

The explanation for mysteries 1 to 3 is straightforward: School children from

---

[14] Note that the term "Asian" as used by the IQ testers, like their use of terms such as "bilingual," is almost too inexplicit to be useful. It is ludicrously broad.

minority language backgrounds have generally been tested in what for them is a nonprimary language (English) on IQ tests, achievement tests, and in all sorts of diagnostic procedures used for placement (Cummins 1984, 1986; Figueroa 1989; Gersten & Woodward 1994; Valdés & Figueroa 1994). In spite of this misguided practice, racial inferences have often been drawn from such testing. Instead, reasonable inferences about English language proficiency should have been drawn. As for Asian children who excel and quickly learn English (mystery 4), examination of the literature will show that in many cases these children are from the high-end of the Asian socio-economic spectrum and have been immersed in mainstream English-speaking communities. Their access to English-language experience, therefore, is more conducive to rapid acquisition of English than for other minority language groups (including some "Asians") living in socio-economically less mobile settings and who have had less access to English (or another mainstream language). If we look at Asian children across the wide socio-economic spectrum, the language-factor theory predicts that they will not differ from other racially or ethnicly defined groups, e.g., see the nearly equivalent primary language scores of similar groups of Chinese and Americans in Xiao and Oller (1994) and between Thais, Vietnamese, and Americans in Oller, Bowen, Dien, and Mason (1972).[15]

A similar straightforward explanation follows for differences favoring whites over blacks (mystery 5). In the groups that have been compared, as the vast quantities of data from the alpha and beta tests show, the Afro-American samples included subjects from all over the country, some who had attended segregated schools in the South and some who had not attended school at all, along with some who had enjoyed greater access to mainstream English dialects (especially in the North — as the testers reluctantly admitted; cf. Yerkes 1921: 351). The reason the mean score on *both* verbal and nonverbal IQ tests for some groups of Afro-Americans is lower than that observed for comparison white groups is owed to differences across dialects for a significant number of the subjects tested (Labov 1970, 1972), but not all of them. The same applies in the case of the Irish and English contrast (mystery 6). The Irish speak a distinct dialect and as Lynn's (1979) data clearly show the differences from the English norm become predictably greater as the tested subjects reside farther from London (see Benson 1995: 226). In fact, the observed difference in test scores for the Irish on English tests is almost exactly the same as that for the Afro-Americans on white-English tests. According to the language-factor theory, some speakers of the minority dialect are more bidialectal in the majority system while others are less so. As a result, since all the tests, verbal and nonverbal alike, depend on the language used by the examiners and in the test items, the differences on both sorts of "IQ" tests are harmlessly predicted.

---

[15] However, evidence that translation of "IQ" tests does not produce equivalence has been argued with empirical data and examples by a number of authors (Manual 1935; Oller 1979: 88-93; Gomez et al. 1983; Figueroa 1989: 147; Valdés and Figueroa 1994: 101-8).

The deaf-hearing contrast can also be accounted for by the language-factor theory and the same holds for the observed contrast between deaf children of deaf parents compared with deaf children of hearing parents. The fact is that deaf children born to deaf parents not only have access to a rich and well developed language (American Sign Language in the United States; Isham & Kamin 1993), but their main interlocutors are adults in the close-knit deaf community (Wilcox & Wilcox 1991). Moreover, the language-factor theory suggests straightforwardly why such fluent signers should excel on the visual and spatial manipulation tasks of so-called "nonverbal" or "performance" tests. The fact is that visual signers practice visuo-spatial transformations all the time in their manual-visual sign systems. In English literacy, however, they are not expected to have any advantage since English is at best a second language for them, and ASL does not yet have a standardized system of writing. The deaf children of hearing parents by contrast have vastly reduced access to English speech (the merits of lip-reading being mostly exaggerated or misunderstood by hearing persons; Lane, Hoffmeister, & Bahan 1996: 99, 213ff.) and their parents, at best, are usually non-native models of signed language systems because the primary language of the hearing parents is invariably a spoken language. So mystery 7 is disposed of.

A more esoteric part of the puzzle about deaf performances on IQ tests is tied to Braden's (1989) reading of Spearman's (1927) little understood hypothesis about *g*-loadings (also Jensen's reading in 1984: 583) and certain predicted contrasts between minority and majority groups (mystery 8). Spearman had in mind black versus white differences, but this hypothesis, if correct, should generalize to essentially every majority/minority contrast so long as the innateness premise and its related meritocracy theory were correct. Spearman's hypothesis is that the mean of the majority group on measures of *g* will be higher than the mean for the minority group and that this difference will vary in positive correlation with the strength of loadings on *g* (for the minority group). However, if the language-factor theory is correct Spearman's hypothesis must always tend to be true with respect to loadings on every *g*-factor that is defined by tests in the majority language/dialect and it must be false with respect to loadings on every *g*-factor that might be defined with respect to tests in the minority language/dialect.

For instance, suppose that *g* is generally dominated by measures of proficiency in the majority language,[16] then, it follows that the measures that load most heavily on that factor (i.e., the tests that are the best measures of proficiency in the majority language) must also discriminate the most between the majority and minority, in which case Spearman's hypothesis is necessarily true — but not for the reasons that Jensen, Braden, and others have supposed. Or, consider the case where *g* is mainly defined by measures associated with the minority language. For the cases examined by Jensen, Braden, and the IQ enthusiasts these would involve tests in any language/dialect other than the majority dialect of American or British English. For instance, suppose the tests defining *g* were all given in American Sign Language. In such a case, differences between the majority and minority groups on tests heavily loaded on that minority language *g*-factor would be both meaningless and uncorrelated with the loadings on *g* precisely because the majority group (the English speakers) have no proficiency worth speaking of in ASL.

---

[16] To see that this is so for the tests that Spearman (1927) championed as measures of *g* we only need to consult his book to find a remarkable variety of language proficiency measures, together with their loadings on *g*. On page 202, he breaks down one of the longer lists into (1) measures of the "relation of likeness" by tests of (a) "opposites" .89, (b) "synonyms" .85, and (c) "classification" .77; (2) measures of "relation of evidence" by (a) "inferences" .74 and (b) "likelihood" .92; (3) "mixed relations" by (a) "analogies" .79, (b) "completion of sentences" .86, (c) "completion of paragraphs" .78, (d) answering "questions" about "passages" .80, and (e) "comprehension of paragraphs" .94; and (4) "memory" by recalling the substance of a "short story"

However, for deaf children of hearing parents, the tests that define *g* for almost any battery of tests will tend to be heavily influenced by the majority language (English) and some correlation is predicted for deaf versus hearing scores correlated with deaf loadings on that *g*-factor. The same holds for tests in white English applied to speakers of black English, and so forth. We must predict that minority group scores on tests of white English will differ from the means attained by native speakers of white English at just about exactly the same levels as the loadings on *g* will differ from each other. Thus, Spearman's hypothesis is accidentally right in some cases and wrong in others, but even when it predicts the observed result it does so for the wrong reasons. Thus, mystery 8 is handled.

That leaves only mystery 9 — the supposed decline in IQ of U.S. immigrants in the 20 years before the Army testing of 1917 and 1918. As Figueroa (1989) and Valdés and Figueroa (1994) point out, the simplest explanation of this alleged decline is the time required for immigrants to acquire English. Those who came to the U.S. 20 years before the IQ testing took place had ample time to acquire English and came to resemble English-speaking Americans. In view of the fact that the Army data shows a steady improvement in scores on the English language tests (alpha and beta) in each five-year increment of time in the country (per Yerkes 1921: 701-4 and Brigham 1923: 94), I believe that this is the correct explanation. It is also the one that the language-factor suggests.

.79. Except for the effects of monoglottosis, there is little else in any of Spearman's descriptions of a wide assortment of tests of *g* that would qualify as anything other than measures of language proficiency (see Spearman 1927: 203-18). He himself repeatedly acknowledges the role played by "symbols (i.e., words)" (ibid.: 209) in what he regards as most central "abstract," "symbolic," or "verbal" thinking" (ibid.: 210). On pages 376-92, he deals with race and essentially accepts the claims of Brigham (1923) and others who have repeatedly argued for racial differences in innate intelligence as supposedly shown in IQ tests. He says the differences held up when people were tested in their own languages (ibid.: 379), but it is difficult to see how such tests are possible across racial groups in view of the fact that across the broad spectrum, race and language are essentially uncorrelated. There is nothing whatever about being Caucasian, for instance, that precludes one's being a native speaker of Mandarin (supposedly an Asian language), and vice versa for every conceivable pairing of language and race. Also see note 14.

Reformation and Reconceptualization

To reform the use of IQ and other school testing procedures, it is essential to start calling "verbal IQ tests" *measures of* ***primary*** *language skills*. When they are applied to persons for whom the language of the test is not the primary language, they are *measures of* ***second*** *(or* ***nonprimary****) language skills*. If these changes are made, the absurdity of calling a normal person who has not yet learned a certain dialect of English "retarded," "learning disordered," "language impaired," "learning disabled," and the like, will be averted. A more valid interpretation could be made about progress in acquiring a certain brand of English, or Spanish, or whatever the language of the test might be. By the same token, according to the language-factor theory, tests formerly called "*nonverbal* IQ tests" are measures of conceptual skills that are accessible only through proficiency in some particular language, usually the test-taker's primary language. Therefore, it would be more appropriate to call them *measures of the* ***primary*** *language abilities of the test-taker* than to continue to call them "nonverbal IQ tests." Furthermore, unless clear verbal instructions can be given in the primary language of test-takers, the results of all "nonverbal" tasks should be regarded with reasonable skepticism. It almost goes without saying that in suggesting these reforms I take it for granted that the usual requirements for test reliability, validity, and practicality will be required of all the tests that are used. No procedure can be exempted from those exacting requirements.

Contrasts between verbal and nonverbal IQ scores also need to be re-evaluated. According to the language-factor theory, where the language of the verbal test is the same as the primary language of the test-taker, a contrast in verbal and nonverbal IQ scores, evidently, should be viewed as a contrast between abstract thought (accessed through the primary language in the "nonverbal" test) and particular skill in using the surface-forms of the primary language (in the "verbal" test). Also, it may be hypothesized (from the language-factor theory) that partialing out whatever variance in "verbal" tests may be owed to peripheral skills in managing surface-forms of the primary language will leave behind variance that should be very nearly equivalent to the largest source of variance in "nonverbal" tests (*ceteris paribus*). However, in cases where the "verbal" test is administered in a language that is not the primary language of the test-takers, the contrast between those scores and scores on a "nonverbal" test should be viewed as a contrast between first and second language proficiency. All these outcomes would constitute direct empirical evidence against the defense of IQ testers who say that their instruments are pure and unbiased tests of innate intelligence. Evidently, such claims are symptomatic of monoglottosis and are mistaken.

Acknowledgements

Thanks are due especially to my brother D. Kimbrough Oller for the riddle about monolingualism, to my son Mark L. Oller for insisting that I write this paper and for critiquing three earlier drafts of it. I also thank Marie Chavez, Mary Anne Chavez-Oller, Jack S. Damico, Donald Fischer, Eduardo Lopez, John L. Omdahl, Rebecca G. Tierney, Anne Wiltshire, and several anonymous (very well-informed and helpful readers for the journal, in addition, of course, to the editors) for their comments on one or more drafts. Although I am sure that the present version was much improved by their insightful comments, whatever errors remain are my own. Readers are welcome to contact me at joller@usl.edu or by phone at (318) 482-6721.

References

Allen, G. E. 1995. Eugenics Comes to America in Jacoby and Glauberman (eds.) *The Bell Curve Debate: History, Documents, Opinions* 441-75. New York: Random House.

American Psychiatric Association. 1980. *Diagnostic and Statistical Manual of Mental Disorders.* (3rd ed.) Washington, DC: American Psychiatric Association.

American Psychiatric Association. 1987. *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed. rev.) Washington, DC: American Psychiatric Association.

American Psychiatric Association. 1994. *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.) Washington, DC: American Psychiatric Association.

Angelis, P. J. 1972. Listening Comprehension and Error Analysis in G. Nickel (ed.) 1972: *Applied Contrastive Linguistics: Proceedings of the Association Internationale de Linguistique Appliquee, Third Congress, Copenhagen.* Heidelberg, Germany: Julius Groos Verlag.

Asher, J. J. 1969. The Total Physical Response Approach to Second Language Learning. *Modern Language Journal* 59: 3-17.

Baca, L. 1984. Theory and Practice in Bilingual/Cross-Cultural Special Education: Major Issues and Implications for Research, Practice, and Policy *ERIC Document Reproduction Service,* No 341 267.

Bachman, L. F. 1990. *Fundamental Considerations in Language Testing.* New York: Oxford University press.

Baratz, J. 1969. A Bidialectal Task for Determining Language Proficiency in Economically Disadvantaged Negro Children. *Child Development* 40: 889-901.

Benson, C. 1995. Ireland's Low IQ: A Critique, in Jacoby and Glauberman (eds.) *The Bell Curve Debate: History, Documents, Opinions* 222-33. New York: Random House.

Bickerton, D. 1981. *Roots of Language.* Ann Arbor, MI: Karoma.

Bickerton, D. 1995. *Language and Human Behavior.* Seattle: University of Washington Press.

Binet, A. 1911. New Investigations upon the Measure of the Intellectual Level among School Children. *L'Annee Psychologique* 11: 145-201.

Binet, A. and T. Simon. 1905. New Methods for the Diagnosis of the Intellectual Level of Subnormals. *L'Annee Psychologique* 5: 191-244.

Binet, A. and Simon, T. 1916. *The Development of Intelligence in Children.* Tr. Elizabeth Kite. Baltimore: Williams and Williams.

Bouchard, T. J., Jr. 1981. Familial Studies of Intelligence: A Review. *Science* 212: 1055-59.

Bouchard, T. J., Jr., D. T. Lykken, M. McGue, N. L. Segal, and A. Tellegen. 1990. Sources of Human Psychological Differences: The Minnesota Study of Twins Reared Apart. *Science* 250: 223-8.

Braden, J. P. 1989. Fact or Artifact? An Empirical Test of Spearman's Hypothesis. *Intelligence* 13: 149-55.

Brigham, C. C. 1923. *A Study of American Intelligence.* Princeton, NJ: Princeton University Press.

Buros, O. K. 1938-1972. *Mental Measurement Yearbook* (7 Volumes) Highland Park, NJ: Gryphon Press.

Burt, C. 1966. The Genetic Determination of Differences in Intelligence: A Study of Monozygotic Twins Reared Together and Apart. *British Journal of Psychology* 57: 137-53.

Burt, C. 1972. The Inheritance of General Intelligence. *American Psychologist* 27: 175-90.

Carroll, J. B. 1961. Fundamental Considerations in Testing for English Proficiency of Foreign Students in *Testing the English Proficiency of Foreign Students* (pp. 31-40). Washington, DC: Center for Applied

Linguistics.

Carroll, J. B. 1972. Defining Language Comprehension: Some Speculations in R. O. Freedle and J. B. Carroll (eds.): *Language Comprehension and the Acquisition of Knowledge*. New York: Wiley.

Carroll, John B. 1976. Psychometric Tests as Cognitive Tasks: A New Structure of Intellect in L. B. Resnick (ed.): *The Nature of Intelligence* (pp. 27-56). Hillsdale, NJ: Lawrence Erlbaum Associates.

Carroll, J. B. 1983. Psychometric Theory and Language Testing in Oller, J. W. (ed.): *Issues in Language Testing Research*, 80-107. Rowley, MA: Newbury House.

Carroll, J. B. 1993. *Factors of Cognitive Ability*. New York: Cambridge University Press.

Carroll, J. B. 1995. Reflections on Stephen Jay Gould's *The Mismeasure of Man* 1981. *Intelligence* 21: 121-34.

Chamberlain, P. and P. Medeiros-Landurand. 1991. Practical Considerations for the Assessment of LEP Students with Special Needs in Hamayan and Damico 1991: 111-56.

Choi, I. C. 1994. Content and Construct Validation of a Criterion-referenced English Proficiency Test. *English Teaching* 48: 311-402.

Choi, I. C. 1995. A Comparability Study on SNUCREPT and TOEIC. *Language Research* 31: 357-86.

Chomsky, N. A. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. A. 1995. Language and Nature. *Mind* 104: 1-61.

Clark, J. L. D. 1972. *Foreign Language Testing: Theory and Practice*. Philadelphia: Center for Curriculum Development.

Clark, J. L. D. 1979. Direct Versus Semi-direct Tests of Speaking Proficiency: Theory and Application in E. J. Briere and F. B. Hinofotis (eds.): *New Concepts in Language Testing*. Washington, DC: Teachers of English to Speakers of Other Languages.

Clark, J. L. D. and A. G. Grognet. 1985. Development and Validation of a Performance-Based Test for ESL Survival Skills in P. C. Hauptman, R. LeBlanc, and M. B. Wesche (eds.): *Second Language Performance Testing*. Ottawa, Canada: University of Ottawa Press.

Cohen, A. D. 1980. *Assessing Language Ability in the Classroom*. Rowley, MA: Newbury House.

Cohen, A. D. 1994. *Assessing Language Ability in the Classroom* (2nd ed.) Boston: Heinle and Heinle.

Cummins, J. 1984. *Bilingualism and Special Education; Issues in Assessment and Pedagogy*. Clevedon, Avon: Multilingual Matters.

Cummins, J. 1986. Empowering Minority Students: A Framework for Intervention. *Harvard Education Review* 56: 8-35.

Cziko, G. A. 1983. Psychometric and Edumetric Approaches to Language Testing in Oller, J. W. (ed.): *Issues in Language Testing Research*, 289-308. Rowley, MA: Newbury House.

Damico, J. S., J. W. Oller Jr., and M. E. Storey. 1983. The Diagnosis of Language Disorders in Bilingual Children: Pragmatic Versus Surface-oriented Criteria. *Journal of Speech and Hearing Disorders* 48: 85-394.

Damico, J. S., M. Smith, and L. E. Augustine. 1996. Multicultural Populations and Language Disorders in M. D. Smith and J. S. Damico (eds.): *Childhood Language Disorders*. New York: Thieme Medical Publishers.

Darwin, C. 1874. *The Descent of Man* (2nd ed.) New York: A. L. Burt Co.

De Berkeley-Wykes, J. 1983. Jigsaw Reading in J. W. Oller, Jr., and P. A. Richard-Amato (eds.): *Methods That Work: A Smorgasbord of Ideas for Language Teachers* (pp. 313-320). Rowley, MA: Newbury House.

de Villiers, P. A. and J. G. de Villiers. 1979. *Early Language*. Cambridge, MA: Cambridge University Press.

Dewey, J. 1916. *Essays in Experimental Logic*. Chicago: University of Chicago Press.

Di Pietro, R. J. 1981. Discourse and Real-life Roles in the ESL Classroom. *TESOL Quarterly* 15: 7-33.

Di Pietro, R. J. 1982. The Open-ended Scenario: A New Approach to Conversation. *TESOL Quarterly* 16: 5-20.

Engelskirchen, A., E. Cottrell, and J. W. Oller Jr. 1981. A Study of Reliability and Validity of The Ilyin Oral Interview in Palmer, Groot, and Trosper (eds.): *The Construct Validation of Tests and Communicative Competence*. Washington, DC: Teachers of English to Speakers of Other Languages.

Eysenck, H. J. 1971. *Race, Intelligence, and Education*. London: Tempel Smith.

Ferguson, J. 1984. The Laboratory of Racism. *New Scientist* 103: 18.

Figueroa, R. A. 1989. Psychological Testing of Linguistic-Minority Students: Knowledge Gaps and Regulations. *Exceptional Children* 56: 145-53.

Finocchiaro, M. 1964. *English as a Second Language: From Theory to Practice*. New York: Regents.

Fraser, S. (ed.) 1995. *The Bell Curve Wars: Race, Intelligence, and the Future of America*. New York: Basic

Books.

Gallaudet Research Institute. 1987. *Factors Predictive of Literacy in Deaf Adolescents with Deaf Parents/Factors Predictive of Literacy in Deaf Adolescents in Total Communication Programs.* Report to the National Institute of Neurological and Communicative Disorders and Stroke: Project No. NIH-NINCDS-83-19. Washington, DC: Gallaudet University.

Galton, Francis. 1869. *Hereditary Genius: An Inquiry into its Laws and Consequences.* London: Macmillan.

Gardner, H. 1983. *Frames of Mind. The Theory of Multiple Intelligences.* New York: Basic Books.

Gardner, H. 1993. *Multiple Intelligences.* New York: Basic Books.

Gardner, H. 1995. Scholarly Brinkmanship in Jacoby and Glauberman (eds.): *The Bell Curve Debate: History, Documents, Opinions.* New York: Random House.

Gartner, A., C. Greer, and F. Reissman. (eds.)1974. *The New Assault on Equality: IQ and Social Stratification.* New York: Harper and Row.

Gersten, R. and J. Woodward. 1994. The Language-Minority Students and Special Education: Issues, Trends, and Paradoxes. *Exceptional Children* 60: 310-22.

Ginsberg, A. 1992. Improving Bilingual Education Programs Through Evaluation in C. Simich-Dudgeon (ed.) 1992: *Proceedings of the Second National Research Symposium on Limited English Proficient Student Issues: Focus on Evaluation and Measurement, Volume 1.* Washington, DC: US Department of Education, Office of Bilingual Education and Minority Language Affairs.

Gomez-Palacio, M., E. R. Padilla, and S. Roll. 1983. *Escala de intelligencia para nivel escolar Wechsler* [Wechsler Intelligence Scale for Children]. Ciudad de Mexico: El Manual Moderno, S. A. de C. V.

Gordon, R. A.1980. Labelling Theory, Mental Retardation, and Public Policy: Larry P. and Other Developments Since 1974 in W. R. Gove (ed.): *The Labelling of Deviance.* Beverly Hills, CA: Sage.

Gordon, R. A. 1984. Digits Backward and the Mercer-Kamin Law: An Empirical Response to Mercer's Treatment of Internal Validity of IQ Tests in Reynolds and Brown (eds.) 1984: 507-86.

Gould, S. J. 1981. *The Mismeasure of Man.* New York: W. W. Norton.

Gould, S. J. 1995. Mismeasure by Any Measure in Jacoby and Glauberman (eds.): *The Bell Curve Debate: History, Documents, Opinions.* New York: Random House.

Grimes, J. and B. Grimes. 1993. *The Ethnologue: Language Family Index.* Dallas, TX: Summer Institute of Linguistics.

Hakuta, K. 1986. *Mirror of Language.* Rowley, MA: Newbury House.

Haller, J. S., Jr. 1971. *Outcasts from Evolution: Scientific Attitudes of Racial Inferiority, 1859-1900.* Urbana, IL: University of Illinois.

Hamayan, E. V. and J. S. Damico. 1991. *Limiting Bias in the Assessment of Bilingual Students.* Austin: PRO-ED.

Hasthorne, C. and P. Weiss. (eds.) 1935. *Collected Papers of Charles Sanders Peirce, Volume V.* Cambridge, MA: Harvard University Press.

Hayes-Brown, Z. 1984. Linguistic and Communicative Assessment of Bilingual Children in C. Rivera (ed.): *Placement Procedures in Bilingual Education: Educational Policy and Issues.* Oxford: Oxford University Press.

Heston, J. B. 1975. *Writing English Language Tests.* London: Longman.

Herrnstein, R. J. 1973. *IQ in the Meritocracy.* Boston: Atlantic-Little Brown.

Herrnstein, R. J. and C. Murray. 1994. *The Bell Curve: Intelligence and Class Structure in American Life.* New York: Free Press.

Hilliard, A. G. III. 1984. IQ Testing as the Emperor's New Clothes: A Critique of Jensen's *Bias in Mental Testing* in Reynolds and Brown (eds.) 1984: 139-69.

Hughes, A. 1989. *Testing for Language Teachers.* Cambridge: Cambridge University Press.

Isham, W. P. and L. J. Kamin. 1993. Blackness, Deafness, IQ and *g*. *Intelligence* 17: 37-46.

Jacoby, R. and N. Glauberman. (eds.) 1995. *The Bell Curve Debate: History, Documents, Opinions.* New York: Random House.

Jensen, A. R. 1969. How Much Can We Boost IQ and Scholastic Achievement? *Harvard Educational Review* 39: 1-123.

Jensen, A. R. 1974. How Biased Are Culture-Loaded Tests. *Genetic Psychology Monographs* 90: 185-244.

Jensen, A. R. 1980. *Bias in Mental Testing.* New York: Free Press.

Jensen, A. R. 1984. Test Bias: Concepts and Criticisms in Reynolds and Brown (eds.) 1984.

Jensen, A. R. 1993. Spearman's Hypothesis Tested with Chronometric Information-Processing Tasks. *Intelligence* 17: 47-77.

Jensen, A. R. 1995. The Differences Are Real in Jacoby and Glauberman (eds.): *The Bell Curve Debate: History, Documents, Opinions*. New York: Random House.

Atendra, A. K. and Rohena-Diaz, E. 1996. Language Assessment of Students Who are Linguistically Diverse: Why a Discrete Approach is Not the Answer. *School Psychology Review* 25: 40-56.

Kamin, L. 1995a. Lies, Damned Lies, and Statistics in Jacoby and Glauberman (eds.): *The Bell Curve Debate: History, Documents, Opinions*. New York: Random House.

Kamin, L. 1995b. The Pioneers of IQ Testing in Jacoby and Glauberman (eds.): *The Bell Curve Debate: History, Documents, Opinions*. New York: Random House.

Keller, H. 1908. *The World I Live In*. New York: Century.

Kelly, C. M. 1965. An Investigation of the Construct Validity of Two Commercially Published Listening Tests. *Speech Monographs* 32: 139-43.

Krashen, D. 1985. *The Input Hypothesis*. Oxford: Pergamon.

Krashen, D. and T. Terrell. 1983. *The Natural Approach Language: Language Acquisition in the Classroom*. San Francisco: Alemany.

Labov, W. 1970. Systematically Misleading Data from Test Questions. *Urban Review* 146-69.

Labov, W. 1972. *Language in the Inner City: Studies in the Black English Vernacular*. Philadelphia: University of Pennsylvania Press.

Lado, R. 1957. *Language Testing*. New York: McGraw Hill.

Lambert, W. E. and R. G. Tucker. 1972. *Bilingual Education of Children: The St. Lambert Experiment*. Rowley, MA: Newbury House.

Lane, H., R. Hoffmeister, and B. Bahan. 1996. *A Journey into the Deaf World*. San Diego, CA: Dawn Sign Press.

Linn, R. L. 1989. *Intelligence: Measurement, theory, and public policy*. Urbana, IL: University of Illinois Press.

Lippmann, W. 1922. The Mental Age of Americans. *The New Republic* (October 25). Also in Jacoby and Glauberman (eds.): *The Bell Curve Debate: History, Documents, Opinions*. New York: Random House. [Reference in text is to the latter.]

Luria, A. R. 1966. *Higher Cortical Functions in Man*. New York: Basic Books.

Luria, A. R. and I. F. Yudovich. 1959. *Speech and the Development of Mental Processes in the Child: An Experimental Investigation*. London: Staples Press.

Lynn, R. 1978. Ethnic and Racial Differences in Intelligence: Ethnic and Racial Comparisons in R. T. Osborne et al. (eds.): *Human Variation: The Biopsychology of Age, Race, Sex*. New York: Academic Press.

Lynn, R. 1979. The Social Ecology of Intelligence in the British Isles. *British Journal of Social and Clinical Psychology* 18: 1-12.

Macnamara, J. 1966. *Bilingualism and Primary Education: A Study of the Irish Experience*. Edinburgh: Edinburgh University Press.

Macnamara, J. 1972. Letter re H. J. Eysenck. *Bulletin of the British Psychological Society* 25: 86-79.

Manuel, H. T. 1935. Spanish and English Editions of the Stanford-Binet in Relation to the Abilities of Mexican Children. *University of Texas Bulletin*. Austin, Texas, No. 3532.

Mercer, J. R. 1973. *Labelling the Retarded*. Berkeley, CA: University of California Press.

Mercer, J. R. 1984. What Is a Racially and Culturally Nondiscriminatory Test in Reynolds and Brown 1984: 293-356.

Mintz, S. W. 1972. Review of *Outcasts from Evolution: Scientific Attitudes of Racial Inferiority, 1859-1900*. Urbana, IL: University of Illinois Press, by J. S. Haller, Jr. *American Scientist* 60: 387.

Natalicio, D. and F. Williams. 1971. *Repetition as an Oral Language Assessment Technique*. Austin, TX: Center for Communication Research.

Oakland, T. D. and Parmelee, R. 1985. Mental Measurement of Minority-Group Children in B. B. Wolman (ed.): *Handbook of Intelligence: Theories, Measurements, and Applications*. Hillsdale, NJ: Erlbaum.

Oller, J. W. Jr. 1970. Dictation as a Device for Testing Foreign Language Proficiency. *UCLA Workpapers in TESL* 4: 37-41.

Oller, J. W., Jr. 1978. The Language Factor in the Evaluation of Bilingual Education in J. E. Alatis (ed.): *International Dimensions of Bilingual Education*. Washington, DC: Georgetown University.

Oller, J. W., Jr. 1979. *Language Tests at School*. London: Longman.

Oller, J. W., Jr. (ed.) 1983. *Issues in Language Testing Research*. Rowley, MA: Newbury House.

Oller, J. W., Jr. 1992. Language Testing Research: Lessons Applied to LEP Students and Programs in C. Simich-Dudgeon (ed.) 1992: *Proceedings of the Second National Research Symposium on Limited*

English Proficient Student Issues: Focus on Evaluation and Measurement, Volume 1. Washington, DC: US Department of Education, Office of Bilingual Education and Minority Language Affairs.

Oller, J. W., Jr. 1993. Reasons Why Some Methods Work in J. W. Oller, Jr. (ed.)1993: *Methods That Work: Ideas for Literacy and Language Teachers.* Boston: Heinle and Heinle.

Oller, J. W., Jr. 1994. Challenged Bilinguals. *NABE News* February 1: 15-18.

Oller, J. W., Jr. 1995. Adding Abstract to Formal and Content Schemata: Results of Recent Work in Peircean Semiotics. *Applied Linguistics* 16: 273-304.

Oller, J. W., Jr. 1996. How Grammatical Relations Are Determined in B. Hoffer (ed.): *Proceedings of the 22nd Annual Meeting of the Linguistic Association of Canada and the United States.* Chapel Hill, NC: LACUS.

Oller, J. W., Jr., J. D. Bowen, T. T. Dien, and V. Mason. 1972. Cloze Tests in English, Thai, and Vietnamese. *Languoge Learning* 22: 1-15.

Oller, J. W., Jr., S. Chesarek, and R. Scott. 1991. *Language and Bilingualism: More Tests of Tests.* London: Bucknell University Press.

Oller, J. W., Jr. and J. S. Damico. 1991. Theoretical Considerations in the Assessment of LEP Students in Hamayan and Damico (eds.): *Limiting Bias in the Assessment of Bilingual Students.* Austin, TX: Pro-ed.

Oller, J. W., Jr. and J. Jonz. (eds.) 1994. *Cloze and Coherence.* London: Bucknell University Press.

Oller, J. W., Jr. and K. Perkins. (eds.) 1978. *Language in Education: Testing the Tests.* Rowley, MA: Newbury House.

Oller, J. W., Jr. and K. Perkins. (eds.) 1980. *Research in Language Testing.* Rowley, MA: Newbury House.

Ortiz, Alba A. 1986. Characteristics of Limited English Proficient Hispanic Students Served in Programs for the Learning Disabled: Implications for Policy and Practice (Part II). *Bilingual Special Education* 4: 1-5.

Ortiz, A. A. 1987. Communication Disorders Among Limited English Proficient Hispanic Students. *Bilingual Special Education Newsletter* 7: 1-7.

Ortiz, A. A. and E. Polyzoi. 1988. Language Assessment of Hispanic Learning Disabled and Speech and Language Handicapped Students: Research in Progress in A. A. Ortiz and B. A. Ramirez (eds.): *Schools and the Culturally Diverse Exceptional Student: Promising Practice and Future Directions.* Reston, VA: Council for Exceptional Children.

Ortiz, A. and J. R. Yates. 1983. Incidence of Exceptionality among Hispanics: Implications for Manpower Planning. *NABE Journal* 7: 41-54.

Palmer, A. 1981. Measurement of Reliability and Validity of Two Picture Description Tests of Oral Communication in Palmer, Groot, and Trosper (eds.): *The Construct Validation of Tests of Communicative Competence.* Washington, DC: Teachers of English to Speakers of Other Languages.

Palmer, A., P. J. M. Groot, and G. A. Trosper. 1981. *The Construct Validation of Tests of Communicative Competence.* Washington, DC: Teachers of English to Speakers of Other Languages.

Peirce, C. S. 1865. The Logic Notebook in M. Fisch, C. J. W. Kloesel, E. C. Moore, D. Roberts, L. A. Ziegler, and N. P. Atkinson (eds.) 1982: *Writings of Charles S. Peirce: A Chronological Edition. Volume 1.* Indianapolis, IN: Indiana University.

Peirce, C. S. 1898. Reasoning and the Logic of Things: The Cambridge Conferences Lectures of 1898 in K. L. Ketner (ed.) 1992. Cambridge, MA: Harvard University Press.

Peirce, Charles S. 1908. A Neglected Argument for the Reality of God. *Hibbert Journal* 7: 90-112. Reprinted in Charles Hartshorne and Paul Weiss (eds.) 1935: *Collected Papers of Charles Sanders Peirce, Volume V.* Cambridge, MA: Harvard University Press.

Piaget, J. 1947. *The Psychology of Intelligence.* Totowa, NJ: Littlefield Adams.

Pike, K. L. 1960. Nucleation. *The Modern Language Journal* 44: 291-5.

Pinker, S. 1994. *The Language Instinct.* New York: William Morrow.

Ramey, C. T. 1992. High-Risk Children and IQ: Altering Intergenerational Patterns. *Intelligence* 16: 239-56.

Rivera, C. and C. Simich. 1981. Issues in the Assessment of Language Proficiency of Language Minority Students. *NABE: The Journal for the Association of Bilingual Education* 6: 19-39.

Robertson, G. J. 1972. Development of the First Group Mental Ability Tests in G. H. Bracht, K. D. Hopkins, and J. C. Stanley (eds.): *Perspectives in Educational and Psychological Measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Rueda, R., R. A. Figueroa, P. Mercado, and D. Cardoza. 1984. *Performance of Hispanic Educable Mentally Retarded, Learning Disabled, and Nonclassified Students on the WISC-RM, SOMPA, and S-KABC* (Final Report — Short-Term Study One). Los Alamitos, CA: Southwestern Regional Laboratory for Educational Research and Development.

Saussure, F. de. 1912. *Cours de Linguistique Generale.* Edited by C. Bally, A. Sechehaye, and A. Riedlinger; translated by W. Baskin (Philosophical Library, 1959; McGraw-Hill (edn.) 1966). New York: McGraw Hill.

Savignon, S. J. 1972. *Communicative Competence: An Experiment in Foreign-Language Teaching.* Philadelphia: Center for Curriculum Development.

Scarcella, R. 1978. Socio-drama for Social Interaction. *TESOL Quarterly* 12: 41-6.

Sedgwick, John. 1995. Inside the Pioneer Fund in Jacoby and Glauberman (eds.): *The Bell Curve Debate: History, Documents, Opinions.* New York: Random House.

Simich-Dudgeon, C. (ed.) 1992. *Proceedings of the Second National Research Symposium on Limited English Proficient Student Issues: Focus on Evaluation and Measurement, Volumes 1-2.* Washington, DC: US Department of Education, Office of Bilingual Education and Minority Language Affairs.

Slobin, D. I. (ed.) 1987. *The Crosslinguistic Study of Language Acquisition: Volume 1. The Data.* Hillsdale, NJ: Erlbaum.

Slobin, D. I. and C. A. Welsh. 1967. Elicited Imitation as a Research Tool in Developmental Psycholinguistics. Paper presented at the Center for Research on Language and Language Behavior, Univesity of Michigan, March. Also in C. A. Ferguson and D. I. Slobin (eds.) 1973: *Studies of Child Language Development.* New York: Holt, Rinehart, and Winston.

Spearman, C. E. 1904. General Intelligence, Objectively Determined and Measured. *American Journal of Psychology* 15: 201-92.

Spearman, C. E. 1927. *The Abilities of Man.* New York: Macmillan.

Stern, S. L. 1980. Drama in Second Language Learning from a Psycholinguistic Perspective. *Language Learning* 30: 77-100.

Sternberg, R. J. 1985. *Human Abilities: An Information Processing Approach.* New York: W. H. Freeman.

Sternberg, R. J. 1996. Myths, Countermyths, and Truths about Intelligence. *Educational Researcher* 25: 11-16.

Stoddard, L. 1922. *The Revolt Against Civilization: The Menace of the Under Man.* New York: C. Scribner's Sons.

Stump, T. 1978. Cloze and Dictation as Predictors of Intelligence and Achievement Scores in Oller and Perkins (eds.): *Language in Education: Testing the Tests.* Rowley, MA: Newbury House.

Swain, M., G. Dumas, and N. Naiman. 1974. Alternatives to Spontaneous Speech: Elicited Translation and Imitation as Indicators of Second Language Competence. *Working Papers in Bilingualism: Special Issue on Language Acquisition Studies* 3: 68-79.

Thelen, E. and L. B. Smith. 1994. *A Dynamic Systems Approach to the Development of Cognition and Action.* Cambridge, MA: MIT Press.

Todd, T. L. and T. R. Levine. 1994. Disentangling Listening and Verbal Recall: Related But Separate Constructs? *Human Communication Research* 21: 103-27.

Valdes, G. and R. A. Figueroa. 1994. *Bilingualism and Testing: A Special Case of Bias.* Norwood, NJ: Ablex.

Valette, R. 1964. The Use of Dictée in the French Language Classroom. *Modern Language Journal* 39: 431-4.

Vocate, D. (ed.) 1987. *The Theory of A. R Luria: Functions of Spoken Language in the Development of Higher Mental Processes.* Hillsdale, NJ: Erlbaum.

Vygotsky, L. S. 1934. *Thought and Language.* (ed. A. Kozulin 1986, 1988) Cambridge, MA: MIT Press.

Wilcox, S. E. and P. Wilcox. 1991: *Learning to See: American Sign Language as a Second Language.* Englewood Cliffs, NJ: Prentice Hall Regents.

Wechsler, D. 1974. *Manual for the Wechsler Intelligence Scale for Children — Revised.* New York: Psychological Corporation.

Wing, L. 1981. Language, Social, and Cognitive Impairments in Autism and Severe Mental Retardation. *Journal of Autism and Developmental Disorders* 11: 31-44.

World Health Organization. 1993. *International Classification of Mental and Behavioral Diseases; Diagnostic Criteria for Research.* (10th ed.) Geneva: World Health Organization.

Xiao, S. and J. W. Oller, Jr. 1994. Can Relatively Perfect Translation between English and Chinese Be Achieved? *Language Testing* 11: 267-89.

Yerkes, R. M. (ed.) 1921. *Psychological Testing in the United States Army: National Academy of Sciences, Volume XV.* Washington, DC: Government Printing Office.

Yoakum, C. S. and R. M. Yerkes. (eds.) 1920. *Army Mental Tests.* New York: Henry Holt and Company.

BEST COPY AVAILABLE

Published tests

Brown-Carlsen Listening Comprehension Test. 1953-1955. J. I. Brown and G. R. Carlsen. New York: Psychological Corporation (Harcourt, Brace, Jovanovich).

Colored Progressive Matrices. 1965. J. C. Raven. United Kingdom: H. K. Lewis and Company.

Culture Fair Intelligence Test. 1933-1973. R. B. Cattell and A. K. S. Cattell. Urbana-Champaign, IL: Institute for Personality and Ability Testing.

Goodenough-Harris Draw a Man Test. 1926. D. B. Harris and F. L. Goodenough. New York: Harcourt, Brace, and World.

Goodenough-Harris Drawing Test. 1963. D. B. Harris and F. L. Goodenough. New York: Psychological Corporation (Harcourt, Brace, Jovanovich).

Kaufman Adolescent and Adult Intelligence Test. 1993. A. S. Kaufman and N. L. Kaufman. Circle Pines, MN: American Guidance Service.

Kaufman Assessment Battery for Children. 1983. A. S. Kaufman and N. L. Kaufman. Circle Pines, MN: American Guidance Service.

Kaufman Brief Intelligence Test 1990. A. S. Kaufman and N. L. Kaufman. Circle Pines, MN: American Guidance Service.

Kaufman Developmental Scale. 1972-1975. H. Kaufman. Wood Dale, IL: Stoelting Co.

Lorge-Thorndike Intelligence Tests. 1954-1966. I. Lorge and R. L. Thorndike. Boston: Houghton Mifflin.

Otis Test of Mental Abilities. 1917-1955. A. Otis. Palo Alto, CA: Stanford University. Also available from Sydney, Australia: Australian Council for Educational Research, Department of Psychology, Macquarie University.

Otis-Lennon Mental Ability Test. 1979. A. S. Otis and R. T. Lennon. New York: Psychological Corporation (Harcourt, Brace, Jovanovich).

Otis-Lennon School Ability Test. 1977-1990. A. S. Otis and R. T. Lennon. New York: Psychological Corporation (Harcourt, Brace, Jovanovich).

Raven Progressive Matrices. 1938-1983. J. C. Raven and J. H. Court. New York: H. K. Lewis and Company distributed by the Psychological Corporation (Harcourt, Brace, Jovanovich).

Sequential Tests of Educational Progress. 1956-1972. Educational Testing Service, Princeton, NJ: Cooperative Test Division.

Standard Progressive Matrices-1986 Restandardization. 1989. J. C. Raven. Sydney, Australia: Australian Council for Educational Research at Macquarie University, Department of Psychology.

Stanford Achievement Test (Ninth Edition). 1923-1996. E. F. Gardner, H. C. Rudman, B. Karlsen, J. Merwin, and the Psychological Corporation. New York: Psychological Corporation (Harcourt, Brace, and Jovanovich).

Stanford-Binet Intelligence Scale. 1916. L. M. Terman and H. Goddard. Boston: Houghton Mifflin.

Stanford-Binet Intelligence Scale-Fourth Edition. 1916-1986. R. L. Thorndike, E. P. Hagen, and J. S. Sattler. Itasca, IL: Riverside Publishing Company (Houghton Mifflin). Australian adaptation available from Hawthorn, Victoria: Australian Council for Educational Research.

Test of English as a Foreign Language. 1964-present. College Entrance Examination Board and Educational Testing Service, Princeton, NJ: Educational Testing Service.

Wechsler Adult Intelligence Scale. 1939-1955. D. Wechsler. New York: Psychological Corporation (Harcourt, Brace, Jovanovich).

Wechsler Adult Intelligence Scale-Revised. 1967-1981. D. Wechsler. New York: Psychological Corporation (Harcourt, Brace, Jovanovich).

Wechsler Intelligence Scale for Children. 1949. D. Wechsler. New York: Psychological Corporation (Harcourt, Brace, Jovanovich).

Wechsler Intelligence Scale for Children-Revised. 1974. D. Wechsler. New York: Psychological Corporation (Harcourt, Brace, Jovanovich).

Wechsler Intelligence Scale for Children-Third Edition. 1971-1992. D. Wechsler. New York: Psychological Corporation (Harcourt, Brace, Jovanovich).

Wechsler Preschool and Primary Scale of Intelligence. 1949-1966. David Wechsler. New York: Psychological Corporation (Harcourt, Brace, Jovanovich).

Wechsler Preschool and Primary Scale of Intelligence-Revised. 1967-1989. David Wechsler. New York: Psychological Corporation (Harcourt, Brace, Jovanovich).

Are Americans Becoming More or Less Alike?
Trends in Race, Class, and Ability Differences in Intelligence

Wendy M. Williams and Stephen J. Ceci, Cornell University

American students' test scores have been slowly but steadily declining for the past half century. Some recent explanations for this decline have focused on dysgenic trends resulting from low-IQ parents outbreeding high-IQ parents. In this article, the authors examined the evidence for dysgenic trends by considering race-, class-, and ability-related changes in intelligence test scores over time. They concluded that (a) racial differences in intelligence decreased from 1973 to 1988 and have remained fairly constant since, (b) intelligence differences between the upper and lower thirds of social class groups have been decreasing slightly since 1932, and (c) Preliminary Scholastic Assessment Test-score differences between the top and bottom quartiles have been relatively stable since 1961. Thus, the authors found no evidence supporting the dysgenic hypothesis. Rather, the combined evidence points to a growing convergence across racial, socioeconomic, and ability-related segments of American society.

It's no longer news: American students' test scores have been slowly but steadily declining to the point that the children of America's major trading partners have overtaken American children. Consider several signs of the decline. Despite some recent gains, U.S. children's achievement test scores in math and science have not kept pace with the scores of European and Asian children. Math scores for 13-year-olds fell from fourth place out of 17 nations in 1971 to 14th place by 1986; similar declines can be charted for 9- and 17-year-olds in virtually all areas except reading (Bronfenbrenner, McClelland, Wethington, Moen, & Ceci, 1996). Other indications of decline include the following: Within a 20-year period, Scholastic Assessment Test (SAT) scores slid by a full standard deviation, from 980 in 1963 to 890 in 1981 (Berliner & Biddle, 1995, Table 2.1). Even the number of students in the highest scoring group (those scoring higher than 700 on the Verbal subtest) declined, despite the greatly expanded number of students who took the test (Hayes, Wolfer, & Wolfe, 1996). It is little wonder that U.S. business leaders complain that the workforce of the future may not be able to cope with the increasingly technological and inherently complex nature of work itself (Hunt, 1995).

In view of the rather precipitous decline in cognitive test performance of American students, one can begin to understand the appeal of a major argument advanced in books such as *The Bell Curve: Intelligence and Class Structure in*

Reprinted from *American Psychologist*, vol. 52, No. 11, November 1997.

American Life (Herrnstein & Murray, 1994). This argument attributes changes in test scores to the net result of a combination of beneficial and baleful forces in society. On the positive side, the forces include large increases in early childhood educational spending, adoption, and better nutrition; on the negative side, the forces include dysgenic pressures resulting from low-IQ parents outbreeding high-IQ parents. At times, the positive pressures are seen as offsetting the negative ones, but the negative pressures (dysgenic trends) are seen as dampening possible gains by forcing Americans to struggle against a "headwind": "Mounting evidence indicates that demographic trends are exerting downward pressure on cognitive ability in the U.S. and that these pressures are strong enough to have social consequences" (Herrnstein & Murray, 1994, p. 341).

In this article, we address a set of interrelated issues having to do with perceptions, both real and imagined, that the intellectual ability of various racial, cultural, and economic groups is diverging. This is not a new argument, of course. Variants of it have been around since before Terman normed the Binet test during the earlier part of this century and reported sizable ethnic differences in IQ (see Gould, 1981). Indeed, Galton (1892) made a similar claim long before the inception of modern methods to measure intelligence. But in its modern instantiation, the dysgenesis-divergence argument takes on the power of sophisticated statistical analyses, validation data from industry and the armed forces, and biotechnological advances in brain imaging and genetic mapping of neurologically relevant sites.

For example, Rushton and Ankney (1996) reviewed evidence for the relationship among IQ, race, and brain size, including recent magnetic resonance imaging volumetric estimations of brain size showing significant racial differences in brain volume. Others have reported that the largest racial differences are found on tests possessing the highest heritabilities (Jensen, 1985). Still others have demonstrated that IQ is correlated with central nerve conductance velocity and oscillation (e.g., Reed & Jensen, 1992, 1993). It is understandable that some commentators have explained changes in specific test scores — especially when they fall more in one group than in others — in terms of sophisticated biological data showing group differences in cranial capacity, heritability, and nerve conductance.

But have test scores been declining this century, in general? And, if they have, is there a growing gap between the scores of the top and bottom segments of American society? In this article, we consider three separate sources of data that bear on these questions.

The Genetic Meritocracy Claim

A spate of books and articles have proffered various arguments for the existence of an ever widening test-score gap between the haves and the have-nots.

Author's note. Correspondence concerning this article should be addressed to Wendy M. Williams or Stephen J. Ceci, Department of Human Development, Cornell University, MVR Hall, Ithaca, NY 14853. Electronic mail may be sent to wmw5@cornell.edu or sjc9@cornell.edu.

This gap is said to be creating pressures for social chaos and economic decline (e.g., Eysenck, 1982; Gottfredson, 1998; Itzkoff, 1989; Lynn, 1991, 1998; Rushton, 1995; Seligman, 1992; Waller, 1971; for rebuttals, see Ceci, 1996; Durlauf, Arrow, & Bowles, 1998; Fischer, Houts, Chodrow, & Duster, 1996; Fraser, 1995; Jacoby & Glauberman, 1995). The most comprehensive of these new books is, of course, *The Bell Curve* (Herrnstein & Murray, 1994), an 840-page tome full of statistical formulas and charts that has sold more than a million copies and that has seized the media's attention.

In *The Bell Curve,* Herrnstein and Murray (1994) argued that a genetically induced bifurcation of intelligence may be taking place as a result of the tendency for the offspring of high-IQ, disproportionately white professionals to attend elite colleges and universities, where they meet and marry the offspring of other high-IQ professionals. According to Herrnstein and Murray, the offspring of these high-IQ pairings are statistically more likely to possess higher IQs than are the offspring of low-IQ pairings. This claim is consistent with the statistical evidence recently reviewed by the American Psychological Association's Task Force on Intelligence (Neisser et al., 1996).[1] This dysgenic pressure is claimed to disproportionately affect the intelligence test scores of blacks, Latinos, and socioeconomically disadvantaged individuals:

> Blacks and Latinos are experiencing even more severe dysgenic pressures than whites, which could lead to further divergence in future generations.... Putting the pieces together, something worth worrying about is happening to the cognitive capital of the country (Herrnstein & Murray, 1994, p. 341).... The effect is dysgenic when a low-IQ group has babies at a younger age than a high-IQ group.... In the United States women of lower intelligence have babies younger than women of higher intelligence (p. 351).... The higher fertility rates of women with low IQs have a larger impact on the black population than on the white. The discrepancies are so dramatically large that the probability of further divergence seems substantial. (pp. 353-354)

Importantly, Herrnstein and Murray (1994) wrote about dysgenic *pressures* rather than dysgenic *effects.* They did so because compensating and countervailing pressures in one direction may be canceled by pressures in the opposite direction. For example, Herrntein and Murray argued that although the tendency for lower IQ persons to have more offspring than higher IQ persons (particularly during times of economic scarcity) may have potentially deleterious consequences for the cognitive capital of America, positive pressures in the environment, such as early childhood and educational interventions, could offset this dysgenic effect and could even produce a net increase in test scores. However,

---

[1] And because low-IQ parents tend to start childbearing at an earlier age than do their high-IQ counterparts, they end up having greater numbers of children across generations (this will be true even if the actual number of offspring remains the same in both groups within a given generation).

Herrnstein and Murray went on to assert that "whatever good things we can accomplish with changes in the environment would be that much more effective if they did not have to fight a demographic headwind" (p. 342).

In this article, we asked whether an intellectual dysgenesis has been taking place and, if it has, whether racial, socioeconomic, and ability-related gaps in intelligence are widening. In the first part of this article, we reviewed the descriptive demographic data. Next, we examined three forms of possible divergence. In the final part of this article, we attempted to determine if the alleged widening of the cognitive gap will continue into the next century, and, if so, we considered the policy implications of such a trend. Throughout this discussion, we focused on test-score declines as the most visible (and testable) form of dysgenesis. However, we acknowledge the possibility that downward genetic pressures could be obscured or offset by positive societal interventions, with the result being that test scores will not decline even while the gene pool becomes poorer. But it is the observable decline of American students' test scores that has animated recent alarm about dysgenesis, and it is at this level of discourse that this article is focused.

Descriptive Data

Throughout this century, whites have outscored blacks and Hispanics on IQ tests as well as standardized achievement tests. The gap most commonly reported is approximately one standard deviation. On the most widely used individual IQ test, the Wechsler series, one standard deviation translates into a 15-point gap between blacks and whites, with Hispanics falling midway between these groups.[2] Providing evidence for this claim, Lynn (1996) reported that a representative sample of 2,260 children between the ages of 6 and 17 years revealed that the average IQs of whites and blacks were 103 and 89, respectively; Peoples, Fagan, and Drotar (1995) reported a similar gap of one standard deviation between 3-year-old white children's IQs ($M = 100$) and black children's IQs ($M = 85$) on the current edition of the Stanford-Binet Intelligence Scale.

Racial and ethnic gaps on IQ and achievement tests have existed throughout this century. For example, IQ differences between blacks and whites were evident on the first Stanford-Binet IQ test normed in 1932, and earlier signs of a racial gap of approximately one standard deviation were apparent on the tests administered to recruits during World War I (Loehlin, Lindsay, & Spuhler, 1975). Differences between the scores of rich and poor samples also have been observed since the first tests were administered. Although none of these facts are in dispute among researchers, their meaning and putative causes are.

---

[2] Although the black-white gap in IQ scores is agreed to be on the order of one standard deviation, there is less consensus on the Asian-white gap in IQ scores, which, summing all studies, is on the order of Asian Americans scoring approximately three points higher than whites. However, there is some debate among researchers on the magnitude of this gap. For example, Vernon (1982) reported a mean IQ of 106 among Asian Americans, and Lynn (1996) reported a mean IQ of 107 among Asian Americans; conversely, Neisser et al. (1996) reported a mean IQ of only 98 for Asian Americans, on the basis of Flynn's 1991 estimate. In addition, Herrnstein and Murray's (1994) own analyses of the National Longitudinal Assessment of Youth data revealed a mean IQ of 106 for Asian Americans. Taken together a prudent position seems to be that IQ scores of Asian Americans are slightly higher than those of whites.

Below, we address the following four questions: (a) What are the achievement test score trends for various racial, socioeconomic, and ability-based groups in America? (b) What do these trends reveal about IQ changes among members of these groups? (c) How can these changes be explained? and (d) What do these data portend for public policy in America's future? Because of the close interrelationship between the first two questions, we considered them in tandem.

Achievement Versus IQ

Readers who are unfamiliar with the psychometric tradition may wonder why we focused on achievement test scores instead of intelligence test scores. After all, the dysgenesis argument is about changes in scores on tests of intellectual aptitude rather than on tests of school achievement. However, putting aside one's theoretical orientation (i.e., whether achievement and aptitude are conceptually different, albeit causally related, or whether they are held to be one and the same entity), the empirical reality is that trends in achievement test scores closely mimic IQ trends (see review in Neisser et al., 1996). Notwithstanding any theoretical distinction one may wish to draw between intelligence and achievement, the empirical reality is that a reliable measure of one is highly correlated with a reliable measure of the other.

For example, in the recent American Psychological Association Task Force on Intelligence, Neisser et al. (1996) noted that a wide range of "content-oriented achievement tests" correlate highly with IQ as well as with all widely used aptitude tests (e.g., the American College Test [ACT], the SAT, the Graduate Record Examination, and the Medical College Admission Test).[3] This explains why intelligence researchers frequently use measures derived from batteries of achievement test scores as a proxy for IQ scores even though at times the content may appear to be quite dissimilar (e.g., some types of IQ tests contain visual matrices, mazes, and puzzles, whereas some achievement test batteries contain only verbal content, such as political editorials or arithmetical word problems). Ceci, Rosenblum, and Kumpf (1998) described this situation as follows:

---

[3] Note that the claim here is not that tests of intelligence and achievement do not differ in principle but merely that it is empirically difficult to disentangle them. Virtually any battery of achievement tests given to a group of individuals (e.g., math, science, verbal reasoning), when factor analyzed, will yield a sizable general factor that correlates highly with those same individuals' IQ scores (also see Cooley & Lohnes's 1976 demonstration that commingled items from a popular mental ability test were difficult to tell apart from items taken from a widely used achievement test). It is often the case that specific achievement test batteries and IQ tests yield factor structures that differ in the magnitude of the first or general factor or in the nature and size of lower order factors (e.g., speed, various so-called crystallized factors, memory). But at the apex of the factor structure in all diverse test batteries is a general factor that permeates most or all subtests that compose the battery, regardless of whether it happens to be an achievement test battery or an IQ test. This general factor is variously termed *fluid intelligence, general intelligence, g, abstract reasoning,* or the *first principal component.* Because it appears to be involved in successful performance on all types of achievement tests as well as IQ tests, it is held responsible for the empirical difficulty of disentangling them.

> If you doubt the above assertion [that trends in IQ test scores mimic achievement test trends], you can do the following experiment: Select a random stratified sample of children and administer to them two batteries of tests, an achievement battery (e.g., mathematics, reading comprehension, scientific reasoning), and the most widely used IQ battery (e.g., the 10 subtests from the Wechsler Adult Intelligence Scale-III [WAIS-III]). Next, distill from the achievement-test-battery intercorrelations a single summative score that captures the covariance among the math, reading, and scientific reasoning scores. (This is traditionally accomplished by taking the first principal component from it.[4]) This summary score of the sample's achievement test scores will be very closely related to the IQ scores for the same sample. Often, the correlations with IQ approach the internal reliability of the IQ itself, i.e., the achievement test summative score correlates with IQ about as highly as IQ correlates with itself (Ceci, 1996). It is rare to find a measure of achievement that does not correlate highly with an IQ score as noted above.

To some, the close relationship between scores on tests that are avowedly designed to measure school achievement (e.g., math, reading, and scientific reasoning) and scores on tests designed to measure so-called intelligence is attributed to the role that the latter is assumed to play in the former. To others, however, the high correlation between achievement and intelligence confirms the folly of creating hard distinctions where none may exist (Cooley & Lohnes, 1976). Whatever one's view on this topic is (we are agnostic), the high degree of correlation between IQ and achievement scores permits the use of one as a proxy for the other. Hence, in this article, we, at times, relied on trends in achievement test scores for hints about correlated trends in IQ scores. At other times, however, we drew directly on trends in IQ scores.

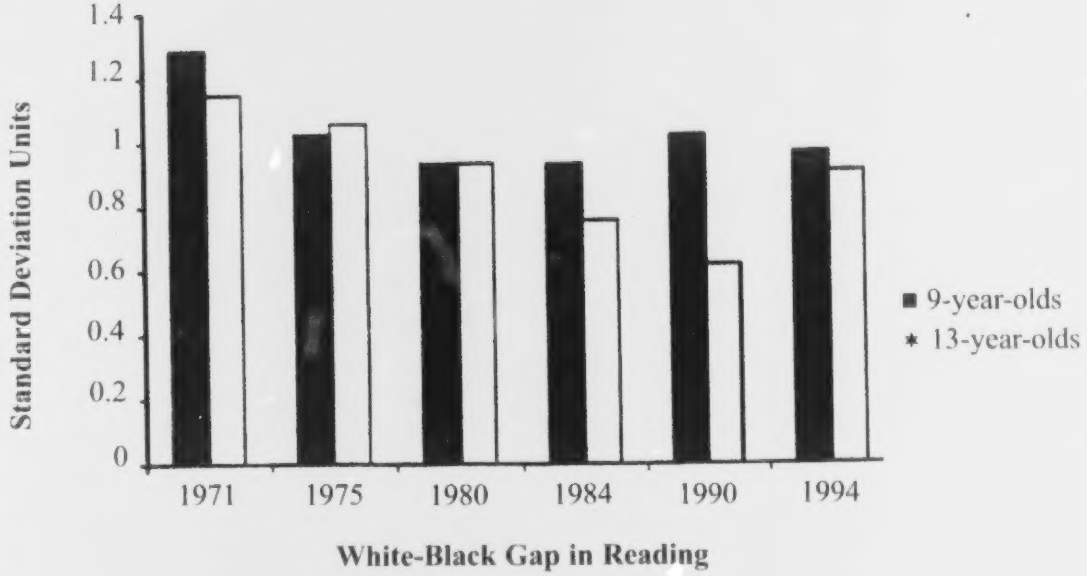
Trends in Racial Differences on Achievement Test Scores

As we already mentioned, one standard deviation (15-16 IQ points) separates the IQ scores of American blacks and whites (e.g., Lynn, 1996; Peoples et al., 1995), and one fifth of a standard deviation (3-4 IQ points) separates the IQ scores of East Asians and whites, the former scoring higher than the latter (Herrnstein & Murray, 1994). On batteries of achievement tests, a similar racial-ethnic gap has been shown to exist. However, recent analyses have shown that the size of this gap narrowed significantly by the late 1980s (Bronfenbrenner et al., 1996; see also Grissmer, Williamson, Kirby, & Berends, 1994; Hauser & Huang, 1998). Although a racial gap of one standard deviation has persistently held for as long as records have been kept, in the early 1970s there began to appear signs that a racial convergence was taking place. Over a period of approximately 15 years, the achievement test scores of black students narrowed the gap with those of white students by between one third and one half, so that by 1986, the racial gap on the National Assessment of Educational Progress (NAEP) had closed by approximately one half (see Figures 1 and 2).

---

[4] This is operationalized as the maximum (linear) variance that can be accounted for, independent of any type of factor rotation, in the matrix of correlations among the various test scores.

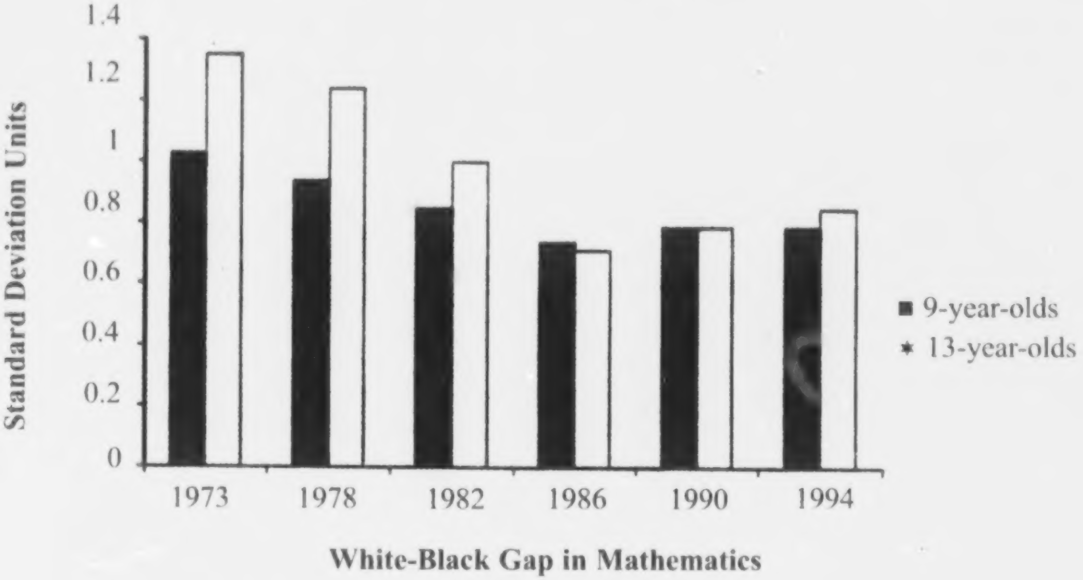
Figure 1: Narrowing of the White-Black Gap in Reading Between 1971 and 1994



There are multiple achievement tests on the NAEP (math, science, reading, and writing), as well as multiple age groups taking these tests (9-, 13,- and 17-year-olds) and multiple grades that crosscut each age group (e.g., on some trend data, 8th graders include not only 13-year-olds but also somewhat older and younger students). Thus, the magnitude of racial convergence varies somewhat by the specific achievement test, grade, and age group being discussed. Between the early 1970s and the late 1980s (National Center for Education Statistics, 1991), substantial gap-closing was evident for nearly all tests, grades, and age groups (for a detailed breakdown, see Hauser & Huang, 1998, Tables 1 and 2).

Unfortunately, however, it appears that the racial convergence in test scores ceased by the late 1980s. The trend to converge did not appear to have continued in the 1994 NAEP trend data that came out in November 1996 (National Center for Education Statistics, 1996). These latest NAEP trend data showed no further systematic changes in racial means, if the discussion is restricted to statistically reliable differences. (After 1988, however, there was a hint of a potential divergence in the scores of blacks and whites, although so far this suggestion of divergence has not reached statistically significant levels.)

Figure 2: Narrowing of the White-Black Gap in Mathematics Between 1973 and 1994



A more detailed analysis revealed the following: In the latest NAEP trend data (National Center for Education Statistics, 1996), there were 12 categories: three age groups (9-, 13-, and 17-year-olds) and four subject areas (science, math, reading, and writing). Comparing the gap between black and white students' scores in these most recent NAEP data with those in the 1990 report revealed that eight of the 12 trends showed a slight divergence between white and black students' test scores but that these differences were not statistically reliable. (In addition, the remaining four of the 12 contrasts showed signs of further convergence, but these also were not significant.) We plotted these data in Figures 1 and 2. Clearly, the next release of NAEP data in 1998 will be extremely interesting: Will the hint of a potential reversal (not statistically significant as of yet) apparent in the 1994 data continue and reach conventional levels of statistical significance? Will there be a reversal of the gains made during the 1970s and the 1980s? These questions are ones to which educators and policymakers should be attuned.

Recently, several independent teams of researchers have reported similar findings through the 1990 NAEP data. Grissmer, Williamson, Kirby, and Berends (1998), using a slightly different statistical model than the one we used, reported highly similar findings in which the 0.9–1.2-standard deviation gap that existed on a broad array of achievement test scores in the early 1970s had been narrowed anywhere from 25 percent to 50 percent by 1986. Along these same lines, Hauser and Huang (1998) reported that the one-standard deviation gap between IQ scores for black and white students that existed in 1973 had narrowed by 1988, on average, by 40 percent. (Hauser and Huang's finding reflects the inclusion of science tests that showed very little change over this period except among 9-year-olds; when only reading and math changes were included, to make their data more comparable with our own as well as with Grissmer et al.'s [1998], the average magnitude of gap-closing among 13- and 17-year-olds was 55 percent.)

In fairness to Herrnstein and Murray (1994), two points should be mentioned: First, growing gaps in the scores of blacks and whites are not necessarily inconsistent with a dysgenesis hypothesis, because countervailing beneficial pressures could be offsetting any negative pressures. However, there is no scientific means of testing the extent to which such claims may be true because they are empirically unfalsifiable. Second, Herrnstein and Murray themselves mentioned three studies showing that the IQ gap in cognitive test performance between blacks and whites seems to be converging. Having made this acknowledgment, however, they went on to raise statistical concerns about one of the studies, and they further noted that in their own analysis of the NAEP data, they found a convergence of math and verbal fluency measures of black and white 17-year-olds but far less than what we, Grissmer et al. (1998), and Hauser and Huang (1998) have reported.[5]

> As the table indicates, black progress in narrowing the test score discrepancy with whites has been substantial on all three tests and across all of the age groups. The overall average gap of .92 standard deviation in the 1969-1973 tests had shrunk to .64 standard deviation by 1990. The gap narrowed because black scores rose, not because white scores fell. Altogether, the NAEP provides an encouraging picture (Herrnstein & Murray, 1994, p. 291).... The question that remains is whether black and white test scores will continue to converge. If all that separates blacks from whites are environmental differences and if fertility patterns for different socioeconomic groups are comparable, there is no reason why they shouldn't. The process would be very slow, however,... reaching equality sometime in the middle of the twenty-first century.... If black fertility is loaded more heavily than white fertility toward low-IQ segments of the population, then at some point convergence may be expected to stop, and the gap could begin to widen again. (p. 293)

Thus, for all nine tests (science, math, and verbal fluency, for each of three age groups), Herrnstein and Murray (1994) agreed that there was racial convergence. However, troubled by the earlier onset of childbearing by black teenagers, these authors cling to a dysgenesis hypothesis. On the basis of the 1990 NAEP analyses reported above, however (see also Grissmer et al., 1998; Hauser & Huang, 1998), if blacks' fertility is loaded downward, it is difficult to explain why all evidence points in the opposite direction from that resulting from a dysgenic trend.

---

[5] Herrnstein and Murray (1994) arrived at a smaller estimate of the size of the convergence in the NAEP data, arguing that approximately 33 percent of the racial gap had been closed by the late 1980s. As Hauser and Huang (1998) correctly noted, however. Herrnstein and Murray arrived at their lower estimate of convergence by using the wrong measure of variance in making their calculations. The appropriate age-corrected (hence smaller within-group) standard deviations led to the higher rates of convergence reported by us, Grissmer et al. (1998), and Hauser and Huang.

If the trends in IQ test scores were to mimic the achievement test score trends just described, one would expect a similar reduction in the historically stubborn racial intelligence-score gap, indexed by the two most widely used IQ tests (i.e., the Stanford-Binet and the Wechsler series). Specifically, black students (who showed the greatest gains on the NAEP achievement test scores in the 1970s and the 1980s) might also exhibit a comparable gain in IQ, thus closing the one-standard deviation gap by approximately one half. It is important to note that closing the racial gap in achievement test scores has resulted from gains in test scores by black students rather than from reductions in white students' scores.

In contrast, because the gains made by blacks ended by 1988, and actually have shown signs of reversing direction since then, it might be expected that the later cohort's IQ scores will show a commensurate gap-widening. Again, this assumes that trends in achievement test scores mimic trends in IQ scores, regardless of the conceptual distinctions one wishes to draw between achievement and intelligence.

As pointed out by demographers and social scientists (e.g., Grissmer et al., 1994; Hauser & Huang, 1998), the NAEP data sets have three unique advantages for those interested in testing the divergence-convergence hypothesis. First, in contrast to the SAT and IQ tests, the NAEP has not changed its item content since its inception in 1969; that is, the very same items are used today that were administered to earlier cohorts, thus enabling a direct comparison of the number answered correctly. Second, the NAEP is administered to a large, representative, national sample of children at each age.

Third, the NAEP sampling procedure does not exclude certain groups (e.g., those not bound for college), as do tests like the SAT. Hence, having shown that the best available data suggest that racial divergence in intelligence is not increasing but actually decreasing, we turn next to an examination of a different form of divergence, namely, that between the test scores of the offspring of rich and poor parents.

Trends in the Intelligence of Different Socioeconomic Status Groups

Recent analyses by the political scientist James Flynn (1998) of New Zealand have called into question the claim of a widening gap in the IQ scores of rich and poor people. The claim that there is a genetically driven socioeconomic status (SES) divergence leads to the expectation of a growing tendency for good genes for IQ to rise to the top of the occupational scale and for bad genes to fall to the bottom. Specifically, this claim leads to the expectation that the IQ gap between the children of the upper and lower income groups has been diverging over time. Flynn was able to provide a strong test of this claim in his most recent analyses.

Flynn's (1998) data came from the Stanford-Binet normative sample tested in 1932 and the Wechsler Intelligence Scale for Children (WISC) samples tested in 1947-1948, 1972, 1985, and 1989. The relevant comparisons involved the mean IQs of children in American families representing a hierarchy of occupations, ranked from professions at the top to unskilled workers at the bottom. Flynn provided details on the specific U.S. census occupational codes used for each IQ test standardization sample as well as the manipulations required to satisfy various measurement issues.[6]

Flynn (1998) measured the difference between the mean IQs of children whose parents were in the top and bottom thirds of occupational status in each standardization period to see whether these groups diverged in IQ over time. As can be seen in Figure 3, for whites, the IQ gap between the top and bottom SES groups was higher than 12 points in 1932, falling to 10 points in 1948. The size of this gap has not changed through the 1989 WISC-III sample. (If the recent Stanford-Binet and WISC-III samples are polled, the gap stands at little more than 9 points.[7])

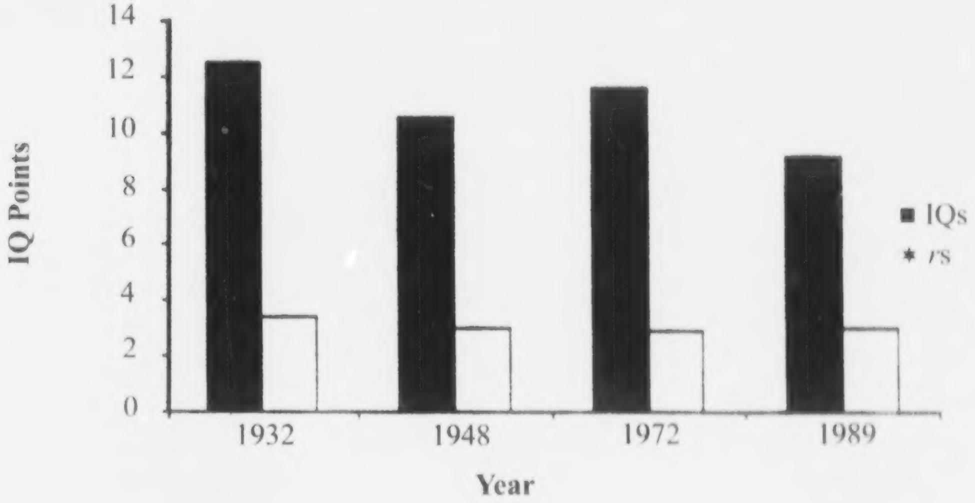
If occupational status is partially determined by genes for intelligence, it would be expected that over time the correlation between occupational status and intelligence would increase as a result of selective mating. As seen in Figure 3, however, the correlation between IQ and occupational status, if anything, actually decreased over time. In summarizing these analyses, Flynn (1998) concluded, "Concerning the effect of social mobility on stratifying genes for intelligence within white America, the most parsimonious conclusion is this: nothing, nothing, nothing, absolutely nothing has happened."

---

[6] Rendering occupational categories comparable across time poses several thorny measurement problems, which Flynn (1998) discussed at length, because the relative ranking of prestige for occupations has changed somewhat over this century. For example, managers were ranked in the second from the top category of occupations in the 1932 Stanford-Binet standardization, but by the 1985 Stanford-Binet and 1989 WISC-III standardizations, managers had been merged with the professions in the top category. Flynn also made adjustments for changes in the way IQ test manufacturers classified the occupational status of American homes: earlier standardization samples were based on the status of the occupation of the male head of household, whereas in 1989, this was based on whichever adult had the highest status occupation. Flynn constructed mean IQs for children whose parents came from each of the five occupational categories to assess whether changes had occurred.

[7] Flynn (1998) showed that the gap for all races combined had increased between 1972 and 1989 from 11.64 to 12.85 points. However, this increase was due solely to increased immigration, plus a natural increase in the number of minorities, which together doubled the number of non-whites in the recent samples. Because non-whites have a lower mean IQ than whites and are concentrated in low-status occupations, such a demographic trend automatically increases the all-races SES-IQ gap. This has nothing to do with the divergence thesis of *The Bell Curve* (Herrnstein & Murray, 1994), which predicts a widening class-IQ gap on the basis of dynamics within groups that have been in America throughout this century but that differ in genetic talent.

Figure 3: IQ Gap Between Upper and Lower Thirds of Social Class



So, on the basis of Flynn's (1998) findings, the claims of social immobility and further widening of the SES gap in intelligence are not supported. Coupled with the racial data reported above, there is reason to doubt the claim of a steady downward pressure on "genes for intelligence" (e.g., Lynn, 1998). Although the magnitude of the narrowing of high- and low-SES groups' IQs is not large, there is absolutely no evidence of an opposite trend or divergence.

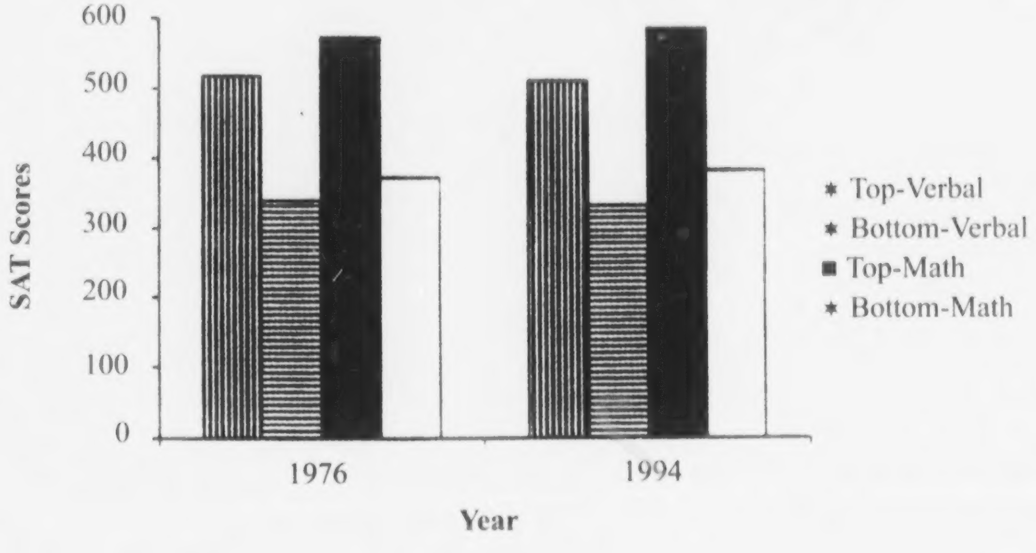
To these two failures to find evidence in support of claims of racial and SES divergences, we add an examination of one final claim of divergence — specifically, the claim that the gap between the scores of the brightest and dullest students is widening. To adumbrate our findings, we found no compelling support for this claim either.

Trends in Intelligence of Ability-Related Groups

Examining changes in average scores between bright and dull students is yet another way of evaluating the dysgenesis-divergence hypothesis. Specifically, if genes for intelligence are leading to increasingly disparate test scores, then it ought to be possible to test this claim by showing that the gap between students at the top and the bottom of the distribution is getting larger. Testing this expectation is problematic because the content of IQ tests (and surrogates like the SAT) changes over time. Even more problematic is the fact that the demographic makeup of the samples taking these tests shifts dramatically over time. One cannot be sure that a widening of the gap between bright and dull test takers over time is due to actual aptitude-related changes in the gene pool over time, as opposed to changes in the proportions of economically disadvantaged students taking these tests (see Berliner & Biddle, 1995; Hayes et al., 1996).

Yet, despite the much larger size of the pool of applicants taking the SAT today as compared with 20 years ago — meaning that less selective, more economically diverse students take the SAT today (Berliner & Biddle, 1995; cf. Hayes et al., 1996) — there is no real evidence of a divergence between the top 10 percent and the bottom 20 percent of scorers on the Verbal and Mathematical sections (see Figure 4). Specifically, the gap between the scores of the highest and lowest groups of students has not widened: In 1976, the mean SAT Verbal score for the top 10 percent of high school seniors was 518, whereas the mean SAT Verbal score for the top 10 percent in 1994 was 512, a decline of 6 points. For the SAT Mathematical section, the comparable scores for the top 10 percent in 1976 and 1994 were 574 and 586, respectively, an increase of 12 points. So, no clear downward trend is evident. But for the purposes of testing the dysgenesis claim, it is necessary to go beyond these data and to ask about the relative gap between the top and bottom groups of students. Here are the relevant means: For the bottom 20 percent of SAT Verbal scores in 1976, the mean score was 339; for the bottom group in 1994, the mean score was 332 — a decline of seven points. In contrast, the means for the SAT Mathematical scores over this same time period were 364 and 363, respectively. Thus, the gap between the SAT scores had not widened significantly over this 18-year period.

Figure 4: Top and Bottom Scoring Scholastic Assessment Test (SAT) Groups for 1976 and 1994



Preliminary Scholastic Assessment Test (PSAT) scores have an advantage over SAT scores because the PSAT has been administered to a nationally representative sample of high school juniors since 1961 (Solomon, 1983), thus avoiding the problems of self-selection that plague interpretations based on the SAT.[8] Berliner and Biddle (1995) showed that there has been no decline in the aggregate trend of PSAT scores, but the question that we ask here is whether the PSAT scores of the top and bottom groups of students have diverged over time. To examine this question, we plotted the gap between the highest and lowest quartiles of juniors from 1961 through the most recent year (1995) for which data are available from the Educational Testing Service.

As can be seen in Figure 5, our analysis suggests that the gap between the top and bottom students has not been widening over the past 35 years. In 1961, the gap in Verbal scores between the top 25 percent and the bottom 25 percent of students was 15.2; this same gap in 1995 was 14.0. Similarly, for the gap in Mathematical scores, the trend was slightly downward going from 15.3 to 15.0 over the pat 35 years.[9] Both correlations between the size of the Verbal and Mathematical gaps over time were negative and significant (*r*s = -.46 and -.43, respectively). This finding confirms what is visibly apparent in Figure 5; namely, the best and worst students have converged on the PSAT over time.

A more direct way to determine whether Americans are diverging is to correlate changes in the size of standard deviations over time, assuming that greater divergence would be reflected in larger standard deviations. However, in the PSAT data, there is evidence that American high school juniors have been converging since 1966 ($r = -.55$ between year of test administration and standard deviation for that year, $p = .002$). (There was no significant correlation between means and standard deviations over this same period [$r = .23$].)

Taken together, these three analyses of alleged divergences (racial, SES, and ability-related) converge on the view that no sizable cognitive dysgenesis seems to have been occurring in America. If there is downward pressure on the gene pool for intelligence, there is no apparent manifestation that we have been able to detect in these three nationally representative data sets (NAEP, WAISIII IQ norms, and PSAT).

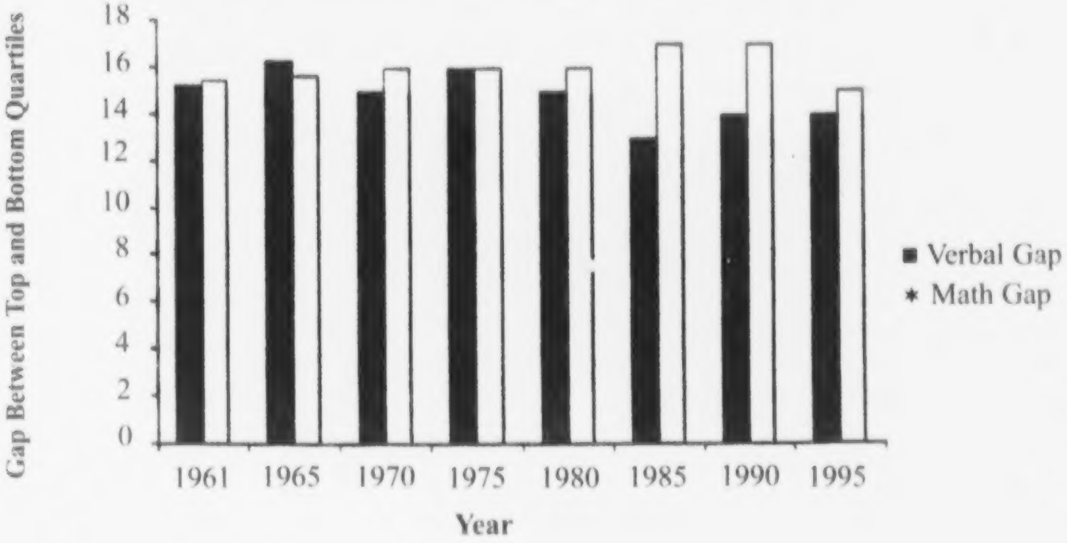
An even more basic problem with the dysgenesis-divergence hypothesis can be raised: If the presumed downward pressure on the gene pool for intelligence is tied to a growing economic stratification, as Herrnstein and Murray (1994)

---

[8] Another interpretative snarl that confronts anyone attempting to make sense of changes in SAT scores is that there have been dramatic shifts in the proportion of colleges that require it; this has resulted in some of the most select students taking the SAT because they wish to attend elite out-of-state colleges because their in-state colleges require the ACT. To whatever degree the distribution of colleges requiring the SAT and the ACT has changed over time, this will confound the interpretation of such results.

[9] We interpolated the 1961–1965 scores, which resulted in some imprecision. Moreover, the 1961 data was based on only 508,000 high school juniors who took the PSAT, whereas by the mid-1960s, the number had jumped to one million. So, some caution is needed in making comparisons of this sort, even though we suspect that any sampling error is in the direction of a less elite sample taking the PSAT over time, thus rendering our agreement even stronger. As support for this position, even if we were to choose later data that are not interpolated but based on large samples that are nationally representative, we would arrive at the same conclusion.

Figure 5: Gap Between the Top and Bottom Quartiles on the Preliminary Scholastic Assessment Test



claimed, how can this claim be reconciled with the empirical reality that society's economic resources are not parceled out in accordance with differences in intelligence? Put bluntly, if income were distributed according to differences in IQ, one would expect a far less asymmetric distribution of income than there is now. "Many more people would earn close to the national mean, and far fewer would earn at either of the extremes" (Ceci & Williams, 1997, p. 1057).

> In a recent econometric analysis, Dickens, Kane, and Schultze (1995) showed that if IQ were equated among all people and only nonintellective variables were allowed to vary (e.g., parental SES and motivation), then the resultant income distribution would resemble the one we now have. Conversely, if all nonintellective differences were equated and income was distributed solely in accordance with differences in IQ scores, then a far more egalitarian income distribution would be observed than the one we now have. (Ceci & Williams, 1997, p. 1057)

Granted, no serious scholar would argue that income is determined solely by IQ; however, Dickens et al.'s analysis demonstrates the hurdles faced by those who maintain that economic variation is due largely to differences in genes for intelligence.

> Another way to think about this is to compare the incomes of those who possess the top 10 percent of IQs with the incomes of those who possess the top 10 percent of wages. The incomes of those with the top 10 percent of IQs in Herrnstein and Murray's (1994) National Longitudinal Survey of Youth sample earned 55 percent more than average-IQ persons earned. In contrast, the top 10 percent of wage earners in this same sample earned 200 percent more than the

> average person earned! Hence, the proportion of the variation in income that can be explained on the basis of variation in IQ is actually rather small. (Ceci & Williams, 1997, p. 1057)

So, the claim by meritocracy advocates that society's resources are being bifurcated as a result of a widening gene pool for intelligence does not mesh with the empirical reality: Income varies much more because of non-IQ differences than because of IQ differences, leading one team of economists to remark, "If all that mattered was [IQ] scores, U.S. society would clearly be very egalitarian. Eliminating differences due to IQ would have little effect on the overall level of inequality" (Dickens et al., 1995, p. 20).

Conclusion

In this article, we have shown that the claims of divergence in Americans' intelligence are not supported by analyses of national data sets of cognitive scores. Although SAT scores did indeed spiral downward in the 1960s and the 1970s, there are good reasons for this decline that appear to have nothing to do with dysgenic trends. For example, the pool of high school seniors taking the SAT became less selective during this period, and the number of universities requiring it increased. Had the SAT been taken by all high school seniors rather than by a self-selected sample in the early days of its administration, we suspect that the decline would have been greatly diminished in magnitude (Berliner & Biddle, 1995).

Support for this assertion was provided by the finding that during the same period for which SAT scores were declining, PSAT scores were increasing. As we have pointed out, PSAT scores are a population-based measure not susceptible to sampling fluctuations due to self-selection. If we contrast the means for a 20-year period during which the size of the pool of juniors taking the test was at least one million, the mean PSAT Verbal score actually rose from 42.7 in 1966 to 48.7 in 1995, and the mean PSAT Mathematical score rose from 45.0 to 48.9 over this same period. This is the best evidence that the often-reported downward trend in SAT scores probably does not reflect any systemic factors.

In addition to the stasis in PSAT scores, there appears to be no divergence in intelligence test scores between upper and lower SES groupings. If anything, there has been approximately a 25 percent convergence in the IQ scores of the upper and lower one thirds of SES groups.

The area of most dramatic convergence (as opposed to divergence) was in racial differences in intelligence during the 1970s and the 1980s. What might account for the finding that black students during this period closed approximately half the gap that had separated them from white students? To answer this question we begin by considering some nongenetic factors that could influence test scores.

As Ceci et al. (1998) and Williams (1998) have pointed out, potential factors responsible for the increase in blacks' test scores include substantial increases in educational spending throughout this entire century (see Hanushek & Rivkin, 1997), especially for programs targeted at minorities (e.g., Head Start; Title 1; and desegregation, busing, and school lunch programs). In addition, there has been an enormous increase in parental educational attainment by black and Hispanic parents during the same period of rapid score gains by black youngsters (Grissmer et al., 1994, 1998).

It is well-known that parental educational level is tied to children's educational attainment (Bronfenbrenner et al., 1996). So, if parents are becoming better educated and if black parents are making disproportionately greater gains in becoming better educated than are white parents, this difference should elevate black students' scores relative to those of other students. Cook and Evans (1996) and others recently have estimated that approximately a quarter of the racial gap was closed as a result of the disproportionately large gains made by black parents in their educational levels during the period in question. Other factors contributing to black students' test gains include a disproportionate reduction in family size for black families over the same time period and the associated increase in financial resources per child that accompanies a reduction in family size.

One could always argue that dysgenic trends are operating but that they are too recent to be detected in the analyses reported in this article. To whatever extent this is the case, there is no way of knowing. The 1996 NAEP data are the most recent that are available, and they also represent the best data source, given their constancy over the past 25 years. In the most recent NAEP trend data, which were released in the fall of 1996, the racial gap appears to have held firm. In general, the gap does not seem to have increased or decreased reliably since the late 1980s. Thus, one could be heartened that the racial gap is not increasing, or one could be saddened that the closing of the gap has not continued. But regardless of one's reaction to these most recent data, they do not support the suspicion that dysgenesis is taking place but rather that it is too recent to be detected.

Along similar lines, one could argue that dysgenesis is occurring but that it cannot be detected by examining changes in mean scores, because differences between top and bottom groups are meaningful only if the standard deviations are the same. Although this is a problem with some data sources, such as IQ, for which test makers force scores into a distribution with a fixed standard deviation, there is evidence in the data sources used in this article that the standard deviations have remained constant over the same period during which means have converged for social classes, ability groups, and races. Because means and standard deviations are independent, this convergence in means constitutes further evidence of a lack of divergence. For example, depending on the age and the test under consideration, blacks' gains on the NAEP have taken place in the face of relatively stable standard deviations (roughly 34 points, collapsing across the three ages and the three content tests). The same is true of the PSAT data: The standard deviations have remained almost identical over the period of rising mean scores.

Finally, how can one reconcile the argument and the analyses reported here with the frequently reported dysgenic effects in the scientific literature (e.g., Vining's 1982 analyses that formed the basis of much of Herrnstein and Murray's 1994 argument)? According to Vining and others (see Lynn, 1998), lower-class and lower educated persons have tended to have larger families than have higher educated persons, and this trend has been evident for nearly 200 years. This is a complex issue, one that goes beyond the constraints of this article. Interested readers can consult the debate between Preston (1998) and Lynn (1998) for their respective views as well as statistical evidence for and against the argument that low-IQ families have been outbreeding high IQ ones. There are other relevant views as well (Loehlin, 1998; Waldman, 1998). Until there is a consensus within the scientific community on this point, there is no alternative to the test-score evidence that has been presented in this article.[10]

In sum, we found no compelling evidence supporting the hypothesis that a dysgenic trend is at work, undermining Americans' intellectual capital. In the battle to control the hearts and the minds of American educational policymakers and opinion shapers in the mass media, it seems imperative that the combatants temper their passions with a close inspection of the data.

---

[10] Although anecdotal, it seems relevant too that many societies throughout the world have far more rigid social structures than the United States, with marriages strictly confined within social strata. Despite centuries of such social immobility, there does not appear to have been any obvious deleterious consequences.

References

Berliner, D. C., & Biddle, B. J. (1995) *The manufactured crisis: Myths, fraud, and the attack on America's public schools.* Reading, MA: Addison-Wesley.

Bronfenbrenner, U., McClelland, P., Wethington, E., Moen, P., & Ceci, S.J. (1996). *The state of Americans: This generation and the next.* New York: Free Press.

Ceci, S. J. (1996). *On intelligence: A bioecological treatise on intellectual development* (Expanded ed.). Cambridge, MA: Harvard University Press.

Ceci, S. J., Rosenblum, T. B., & Kumpf, M. (1998). The shrinking gap between high- and low-scoring groups: Current trends and possible causes. In U. Neisser (Ed.), *Intelligence on the rise: Long-term gains in IQ and related measures.* Washington, DC: American Psychological Association.

Ceci, S. J., & Williams, W. M. (1997). Schooling, intelligence, and income. *American Psychologist, 52,* 1051-1058.

Cook, M.D., & Evans, W. N. (1996, July). *Families or schools? Explaining the convergence in White and Black academic performance.* (Working paper). Monterey, CA: Naval Postgraduate School.

Cooley, W. W., & Lohnes, P. R. (1976). *Evaluation research in education.* New York: Irvington Press.

Dickens, W. T., Kane, T. J., & Schultze, C. L. (1995, Summer). Ring true? A closer look at a grim portrait of American society. *The Brookings Review, 13,* 18-23.

Durlauf, S., Arrow, K., & Bowles, S. (Eds.). (1998). *Meritocracy and equality.* Princeton, NJ: Princeton University Press.

Eysenck, H. J. (1982). *A model for intelligence.* New York: Springer-Verlag.

Fischer, C., Houts, M., Chodrow, N., & Duster, T. (1996). *Inequality by design: Cracking the myth of the bell curve.* Princeton, NJ: Princeton University Press.

Flynn, J. R. (1991). *Asian-Americans: Achievement beyond IQ.* Hillsdale, NJ: Erlbaum.

Flynn, J. (1998). IQ trends over time: Intelligence, race, and meritocracy. In S. Durlauf, K. Arrow, & S. Bowles (Eds.), *Meritocracy and equality.* Princeton, NJ: Princeton University Press.

Fraser, S. (1995). *The bell curve wars.* New York: Basic Books.

Galton, F. (1892). *Hereditary genius.* London: Macmillan.

Gottfredson, L. S. (1998). Why *g* matters: The complexity of everyday life. *Intelligence.*

Gould, S. J. (1981). *The mismeasure of man.* New York: Norton.

Grissmer, D. W., Williamson, S., Kirby, S. N., & Berends, M. (1994). *Student achievement and the changing American family.* Washington, DC: RAND Institute on Education and Training.

Grissmer, D. W., Williamson, S., Kirby, S. N., & Berends, M. (1998). Exploring the rapid rise in Black achievement scores in the United States (1970-1990). In U. Neisser (Ed.), *Intelligence on the rise: Long-term gains in IQ and related measures.* Washington, DC: American Psychological Association.

Hanushek, E. A., & Rivkin, S. G. (1997). Understanding the twentieth century growth in U.S. school spending. *Journal of Human Resources, 32,* 35-68.

Hauser, R. M., & Huang, M: H. (1998). Trends in Black-White score differentials. In U. Neisser (Ed.), *Intelligence on the rise: Longterm gains in IQ and related measures.* Washington, DC: American Psychological Association.

Hayes, D. P, Wolfer, L. T, & Wolfe, M. F. (1996). Schoolbook simplification and its relation to the decline in SAT Verbal scores. *American Educational Research Journal, 33,* 1-18.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life.* New York: Free Press.

Hunt, E. (1995). *Will we be smart enough?* New York: Sage.

Itzkoff, S. W. (1989). *The making of the civilized mind.* New York: Longmans.

Jacoby, R., & Glauberman, N. (1995). *The bell curve debate.* New York: Random House.

Jensen, A. R. (1985). The nature of Black-White differences on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences, 8,* 193-263.

Loehlin, J. C. (1998). Whither dysgenics? Comments on Lynn and Preston. In U. Neisser (Ed.), *Intelligence on the rise: Long-term gains in IQ and related measures.* Washington, DC: American Psychological Association.

Loehlin, J. C., Lindsay, G., & Spuhler, J. (1975). *Race differences in intelligence.* San Francisco: Freeman.

Lynn, R. (1991). Race differences in intelligence: A global perspective. *Mankind Quarterly, 31,* 255-296.

Lynn, R. (1996). Racial and ethnic differences in intelligence in the U.S. on the Differential Ability Scale. *Personality and Individual Differences, 20,* 271-273.

Lynn, R. (1998). The decline of genotypic intelligence. In U. Neisser (Ed.), *Intelligence on the rise: Long-term gains in IQ and related measures.* Washington, DC: American Psychological Association.

National Center for Education Statistics. (1991). *National Assessment of Educational Progress (NAEP). Trends in academic progress, 1970–1990.* Washington, DC: U.S. Department of Education.

National Center for Education Statistics. (1996). *National Assessment of Educational Progress (NAEP), 1994 long-term trend assessment.* Washington, DC: U.S. Department of Education.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77-101.

Peoples, C. E., Fagan, J. R, III, & Drotar, D. (1995). The influence of race on 3-year-old children's performance on the Stanford-Binet: Fourth Edition. *Intelligence, 21,* 69-82.

Preston, S. H. (1998). Differential fertility by IQ and the IQ distribution of a population. In U. Neisser (Ed.), *Intelligence on the rise: Long-term gains in IQ and related measures.* Washington, DC: American Psychological Association.

Reed, T. E., & Jensen, A. R. (1992). Conduction velocity in a brain nerve pathway of normal adults correlates with intelligence level. *Intelligence, 16,* 259-272.

Reed, T. E., & Jensen, A. R. (1993). Choice reaction time and visual pathway conduction velocity both correlate with intelligence but appear not to correlate with each other: Implications for information processing. *Intelligence, 17,* 191-203.

Rushton, J. P. (1995). *Race, evolution, and behavior: A life history perspective.* New Brunswick, NJ: Transaction.

Rushton, J. P., & Ankney, C. D. (1996). Brain size and cognitive ability: Correlations with age, sex, social class, and race. *Psychonomic Bulletin and Review. 3,* 21-36.

Seligman, D. (1992). *A question of intelligence: The IQ debate in America.* New York: Birch Lane Press.

Solomon, R. J. (1983). *Information concerning the mean test scores for the GMAT, GRE, LSAT, PSAT and SAT for the National Commission on Excellence in Education.* Princeton, NJ: Educational Testing Service.

Vernon, P. E. (1982). *The abilities and achievements of Orientals in North America.* New York: Academic Press.

Vining, D. R. (1982). On the possibility of reemergence of a dysgenic trend with respect to intelligence in American fertility differentials. *Intelligence, 6.* 241-264.

Waldman, I. D. (1998). Complexities in inferring dysgenic trends for intelligence. In U. Neisser (Ed.), *Intelligence on the rise: Longterm gains in IQ and related measures.* Washington, DC: American Psychological Association.

Waller, J. H. (1971). Achievement and social mobility: Relationships among IQ score, education, and occupation in two generations. *Social Biology, 18,* 252-259.

Williams, W. M. (1998). Are we raising smarter children today? School- and home-related influences on IQ. In U. Neisser (Ed.), *Intelligence on the rise: Long-term gains in IQ and related measures.* Washington, DC: American Psychological Association.

Self-Report Measures of Intelligence:
Are They Useful as Proxy IQ Tests?

Delray L. Paulhus, Dana C. Lysy, and Michelle S. M. Yik,
University of British Columbia

Correlations between single-item self-reports of intelligence and IQ scores are rather low (.20-.25) in college samples. The literature suggested that self-reports could be improved by three strategies: (1) aggregation, (2) item weighting, and (3) use of indirect, rather than direct, questions. To evaluate these strategies, we compared the validity of aggregated and unaggregated versions of direct measures with four indirect measures (Gough's Intellectual efficiency scale, Hogan's Intellect composite scale, Steinberg's Behavior Check List, and Trapnell's Smart scale). All measures were administered to two large samples of undergraduates (Ns = 310, 326), who also took an IQ test. Although results showed some success for both direct and indirect measures, the failure of their validities to exceed .30 impugns their utility as IQ proxies in competitive college samples. The content of the most valid items referred to global mental abilities or reading involvement. Aggregation benefited indirect more than direct measures, but prototype-weighting contributed little.

Can people validly rate their own intelligence? Skeptics argue that such self-reports are hopelessly contaminated with a variety of distortions including self-deception, impression management, and reconstrual. Such defensive reactions may explain the low variance in self-ratings of intelligence: Rarely do people rate themselves as "below average" (McCrae, 1990). Even the most forthright and insightful individuals, skeptics warn, can never confirm the veracity of their self-assessments because the concept itself is so elusive in nature.

Despite its elusiveness, the concept of intelligence plays a central role in psychological research, particularly in such contexts as educational evaluation, personnel selection, and child development. To facilitate such research, considerable effort has been devoted to developing self-report alternatives to cumbersome IQ tests of intelligence. Such approaches have progressed well beyond a simple request to "rate how intelligent you are." Three strategies, in particular, have been recommended. One is the use of indirect assessment to bypass the inevitable defensiveness of a direct request for a self-rating. A second, the aggregation strategy, favors multiple-item over single-item measures. A third strategy, prototype weighting, takes into account the differential importance of items within a measure. In this report we evaluate these three strategies by

---

Reprinted from *Journal of Personality* 66:4, August 1998.

determining their ability to improve prediction of performance on IQ tests.

Use of IQ tests as the criterion for intelligence ratings has not yielded high validities,[1] particularly in college samples. The validities are somewhat higher for observer-ratings than for self-ratings. Values in the range of .25 to .50 have been found when the judgment is made by spouses (Bailey & Mettetal, 1977), by friends and strangers (Borkenau, 1993), by adolescent acquaintances (Bailey & Hatch, 1979), and by long-term discussion-group colleagues (Paulhus & Morgan, 1997). Except in the case of spouses, however, the achievement of these solid validities required aggregation across multiple observers.

Self-perceptions typically parallel other-perceptions but, to the extent that the trait being evaluated is highly evaluative (e.g., intelligence), the former are noticeably less valid (John & Robins, 1993). Studies using IQ test scores as the criterion have yielded single-item validities of .32 (Borkenau & Liebler, 1993) and .38 (Reynolds & Gifford, 1996) in general population samples. But in college samples the validities never exceed .30: for example, .26 (DeNisi & Shaw, 1977), .25 (Paulhus & Morgan, 1997), and .26 (Reilly & Mulhearn, 1995). This modest level of validity is the starting point for the present report.

Improving Self-Report Measures of IQ

The literature contains a number of self-report instruments that show potential as proxy IQ scales, that is, economical substitutes for IQ tests. If valid, such scales have great practical advantages over their traditional counterparts: rather than running subjects one-by-one in a tightly supervised laboratory setting, researchers can administer such scales quickly to large groups of subjects. Moreover, self-report questionnaires are less threatening than IQ tests and therefore more likely to elicit cooperation.

Of course, such advantages are pointless unless validities can be improved over those values cited above. An ideal proxy scale would represent, in effect, a parallel measure showing a validity equal to the reliability of IQ tests, that is, upwards of .90. Given that standard IQ tests differ somewhat in emphasis, however, a more appropriate upper limit is the correlation between two well-validated IQ tests, that is, roughly, .80–.85 (Thorndike,1982). Even that level of validity seems unlikely for proxy tests, given that they emphasize typical performance as opposed to the maximal performance tapped by IQ tests (Ackerman & Heggestad, 1997; Paulhus & Martin, 1987).

Potential proxy scales in the literature have relied on three strategies for improving the validity of self-report measures of intelligence. First is the reduction of evaluation-threat by using subtle, nonobvious questions. We use the term "indirect measures" to describe test formats that mask the purpose of the test. The second strategy involves aggregating a set of items to improve reliability.

---

[1] The term "validity" is used to mean correlation with a specific criterion. Its use does not imply that IQ is the sole criterion for measuring intelligence.

The third strategy is weighting the items according to their importance.

Indirect Measurement Strategy

Rather than referring directly to intelligence, items on indirect measures concern interests, behaviors, personality, and the like. In this report, we examined four such measures: Gough's Intellectual efficiency (Ie) scale, Hogan's Intellect composite scale, Sternberg's Behavior Check List (BCL), and Trapnell's Smart scale. All four have shown some validity in predicting criterion measures of intelligence. Although all have an indirect format, the rationale for each is rather different.

Gough's Intellectual efficiency (Ie) scale. In the first such effort, Gough (1953) developed a set of self-report items for use as a proxy measure of intelligence. He administered a pool of items assumed to tap aspects of personality associated with intelligence. Those 52 items correlating most highly with an IQ test in a sample of high school students were assembled and labeled the Intellectual efficiency (Ie) scale. In four cross-validation studies, Gough reported a mean validity of .47. In the four educated samples reported in the latest manual, however, the median validity is only .29 (Gough, 1996).

As for most tests derived from contrasted groups, the Ie items are rather heterogeneous; topics included self-confidence, neuroticism, and social skills as well as intellectual abilities and interests. The vast majority were subtle indicators, that is, they lacked face-validity as indicators of intelligence. As such they are less likely to trigger self-presentation.

Hogan's Intellect composite. Welsh (1975) developed the notion of "intellectance" to denote the "cognitive and interpersonal style that causes people to be perceived as bright." Hogan and Hogan (1992, p. 12) followed this peer-perception notion of intellect in assembling a set of items. A factor analysis revealed two factors. One was labeled Intellectance: "the degree to which a person is perceived as bright, creative, and interested in intellectual matters." The other factor was labeled School Success: "the degree to which a person seems to enjoy academic activities and to value educational achievement for its own sake."

Intellectance items refer to science ability, curiosity (about the world), thrill seeking, interest in intellectual games and generating ideas (ideational fluency), and interest in culture items, while School Success items concern education (being a good student), math ability, good memory, and enjoyment of reading. Observers tend to see high scorers on the Intellectance scale as "imaginative, inventive, and quick-witted, but easily bored and inattentive to detail" whereas low scorers tend to be "unimaginative, narrow, tolerant of boredom, and not needing much stimulation." In contrast, high scorers on the School Success scale are seen as "foresighted, thorough, and painstaking," whereas low scorers are seen as "touchy, restless, and impulsive" (Hogan & Hogan, 1992, p. 40).

Sternberg's Behavior Check List (BCL).[2] *As* part of his investigation into conceptions of intelligence, Sternberg (1988) developed the Behavioral Check List (BCL), a list of 41 behaviors that lay judges associated with intelligence (p. 238). Factor analyses indicated three clusters of items labeled Practical Problem Solving (PS), Verbal Ability (VA), and Social Competence (SC). In a community sample, correlations of the full-scale BCL with an IQ test were found to be .24 with unit-weighting of all items, and .52 when items were weighted according to diagnosticity. All these results were later replicated by Cornelius, Kenny, and Caspi (1989).

Sternberg recommended the BCL as a valuable supplementary measure of intelligence for a number of reasons. Compared to providing a global assessment of their ability, subjects should feel less-threatened by rating specific behaviors and, accordingly, be more accurate. A composite of a large set of these specific behaviors could then yield a maximally valid self-report. Finally, the BCL coverage was designed to extend beyond those aspects of intelligence measured by IQ tests (Sternberg, 1988).

Trapnell's Smart scale. The four-item Smart scale assesses self-appraised intelligence indirectly via statements about the respondent's social reputation (Trapnell, 1994). The content of three of the items was based on the assumption that range restriction in self-ratings due to desirable responding can be reduced by the use of extreme qualifiers (e.g., "very" "extremely," "exceptionally") and by shifting the implied locus of evaluation from the self to others (e.g., "I'm considered to be . . ." in place of "I am. . ."). A fourth item assessed self-reported school grades, based on the assumption that grades provide an indirect but objective index of mental ability that can be recalled and self-reported fairly accurately. The Smart scale correlated .33 with an IQ test in a college sample (Trapnell & Scratchley, 1996).

Aggregation Strategy

Aggregation is a widely accepted strategy for improving the reliability (and therefore the predictive validity) of a measure by decreasing its error of measurement (for a strong case, see Epstein, 1983). Other things being equal, the addition of items similar to those already included should increase the validity of a self-report intelligence measure. In fact, the amount of improvement in reliability and validity can be estimated with available prophecy formulas (e.g., Gulliksen,1967).

The reader may have noted, however, that the number of items in the indirect measures reviewed above varies dramatically (from 4 to 52). This natural variation should allow us to evaluate the utility of aggregation between-measures as well as within-measures.

---

[2] Note that this instrument is not a checklist in the strict sense of requiring respondents to check off answers. Instead, the items are rated in a Likert format.

Weighting Strategy

Typically, aggregation is performed simply by summing or averaging the available items: That is, equal (unit) weights are applied to all items. Although psychometricians usually recommend such unit weights (see Thorndike, 1982), weighting of items by their importance remains an appealing strategy.

One variant of importance weighting was exploited by Sternberg, Conway, Ketron, and Bernstein (1981) in order to improve the validity of the Behavior Check List (p. 49). Their technique, which we will label "the prototype weighting procedure," involved developing a set of weights corresponding to the diagnosticity of each item. A set of judges was asked to rate each BCL item in terms of "how characteristic it is of an ideally intelligent person" (p. 42). Instead of just adding up a respondent's responses to yield a total score, the responses were correlated with the corresponding diagnosticity ratings.

Note that a Pearson correlation is simply the mean of the item-by-item products of the two standardized variables. Therefore this procedure is equivalent to standardizing the 41 weights, multiplying them by the respondent's 41 standardized (within-subject) responses, and averaging the 41 products. The average then represents the subject's composite score. [3] Application of this weighting system by Sternberg et al. (1981) boosted the validity of the BCL up to the .50 range (values of .42-.46 were obtained by Cornelius, Kenny, & Caspi [1989]). Sternberg and colleagues concluded that "a good estimate of IQ can be obtained, based on correspondence between a person's self-perceived pattern of behaviors and the pattern of behaviors in an ideal person" (p. 50).

The Present Study

The self-report measures examined in this report differ with respect to directness (direct vs. indirect) and aggregation (single-item vs. composites). Thus each falls into one of the four categories of a 2 x 2 table (see Table 1). The first category — single-item direct measures — is represented by the adjective "intelligent" (Sample 1) and "clear-thinking, intelligent" (Sample 2). The composite direct measure combined a set of four conceptually similar items. The four indirect measures were Gough's Ie, Hogan's Intellect, Steinberg's BCL, and Trapnell's Smart scale. To evaluate the fourth category of measures — single-item indirect — we calculated the average item validity for each indirect measure. Finally, we compared weighted and unweighted versions of the BCL.

Within the composite indirect category, we also had a specific interest in the comparative validity of the four measures. Although there is some evidence for the validity of each scale, they have never been pitted against one another in

---

[3] A similar approach has been used for some time with Q-sort data where a subject's self-sort is correlated with a particular criterion sort (e.g., Block, 1961, 1971). Again, a higher score indicates that the respondent assigned his/her highest ratings to items that were weighted the highest.

Table 1

A Two-Factor Taxonomy of Sell -Report Measures of Intelligence

	Aggregation Strategy	
	Single Item	**Aggregated Items**
Directness Srategy		
Direct	Global rating of intelligence	Global rating of intelligence plus similar items
Indirect	Average item from an indirect measure	Full indirect measures (BCL, Ic, Intellect, Smart)

predicting a common criterion. A comparative validity study by nonpartisan researchers should provide much more convincing evidence than that offered by the authors of individual scales.

All measures were administered to two large and diverse samples of undergraduates. The criterion for validity of the self-reports was the Wonderlic Personnel Test, a 12-minute IQ test that compares favorably with longer IQ tests. Our analyses focused on comparing the validities of the four categories of measures via correlation and regression techniques. To put all these validities in perspective, we also estimated the performance of ideal proxy scales.[4]

METHOD

Subjects and Procedure

Data were collected from a total of 636 undergraduate students at the University of British Columbia. Sample 1 comprised 310 students (95 males; 208 females; 7 did not specify their gender) enrolled in an introductory psychology course. Sample 2 comprised 326 students (87 males; 205 females; 34 did not specify) enrolled in a second year social-personality psychology course. Approximately 55 percent of the two samples were liberal arts majors, 20 percent science or engineering majors, and 15 percent business majors. All participated for extra marks.

For both samples, subjects were first asked to complete a self-report inventory in group sessions. It included all the direct self-ratings of intelligence. Later, a set of indirect measures of intelligence was distributed in a take-home package, which subjects were asked to complete privately and return for experimental credits. Finally, the IQ test was administered in a separate, supervised session.

---

[4] We did not include factor five measures such as Goldberg's Intellect Scale because they were targeted at personality, not intelligence.

Instruments

Direct measures. A number of intelligence-related items were included in the context of a larger personality inventory. They were selected a priori for their conceptual relevance to intelligence. In Sample 1 they included the following four items: "Is intelligent"; "Is ingenious, a deep thinker"; "Is smart"; and "Is not exceptionally gifted at academic things" (Reverse coded). In Sample 2 the direct items included: "Is clear-thinking, intelligent"; "Wants things to be simple and clear-cut"; "Is clever, sharp-witted"; and "Enjoys thinking about complicated problems."[5] Subjects were asked to rate their agreement with these items on a scale ranging from "1" ("Disagree strongly") to "5" ("Agree strongly"). For special consideration we identified the most face-valid item ("Is intelligent" in Sample 1; "Is clear-thinking, intelligent" in Sample 2).To evaluate the utility of aggregating items, we combined the four items judged by three raters to be the most face-valid indicators of intelligence. Although the scale items differed somewhat in the two samples, the similarity of correlates (see Results section) suggests that the two direct composites measured a similar construct.

Indirect measures. Given our review of the four indirect measures in the Introduction, we will provide only a basic description here. The Intellectual efficiency (Ie) scale of the California Psychological Inventory (Gough, 1953) included 52 True-False statements. The content includes personality-related items, ranging in content from beliefs (e.g., "Success is a matter of will power") and interests (e.g., "I like to read about history") to bizarre items about experiences (e.g., "I have never seen a vision").

Portions of two subscales of the Hogan Personality Inventory (Hogan & Hogan, 1992) were also included to represent the Intellect factor. All items are in True-False format and were developed from a peer-perception view of intellect. For reasons of convenience and space, we limited our selection to 15 items from the Intellectance subscale and 7 items from the School Success subscale. Examples are "I'm good at inventing games, stories, and rhymes" from the Intellectance subscale, and "As a child I was always reading" from the School Success subscale [6]

The Sternberg BCL consists of short, specific, behavioral descriptions originally selected by lay judges as prototypical of intelligent people (Sternberg et al., 1981). We used the final 41-item version provided by Sternberg (1988, pp. 238-239). Our subjects rated from "1" (low) to "9" (high) the extent to which each item was an "accurate self-description." The BCL includes three subscales: the 13-item Verbal Ability subscale (e.g., "Speaks clearly and articulately"), the 15-item Problem Solving subscale (e.g., "Makes good decisions"), and the 13-item Social Competence subscale (e.g., "Responds thoughtfully to others' ideas").

---

[5] Despite its conceptual relevance, we did not include the item "ingenious" in Sample 2 because of confusion the item caused. Apparently, some of our subjects thought the item meant "not a genius."

[6] Two of the Intellect-School Success items are identical to two of the Intellectual efficiency items: We only included them once in our inventory.

The four-item Smart scale measures self-appraised intelligence via simple trait descriptive statements of high face validity (Trapnell, 1994). As with the BCL, subjects rated from "1" (low) to "9" (high) the extent to which each item was an "accurate self-description." The items are: (1) "I'm considered exceptionally or unusually intelligent"; (2) "I'm considered a very 'brainy,' scholarly person"; (3) "I'm considered extremely 'gifted' or talented at academic things"; and (4) "My school grades have usually been near the top of every class."

Objective measure (IQ test). The 12-minute Wonderlic Personnel Test was chosen to assess IQ (Wonderlic, 1992). It is a short-form test of general cognitive ability, that is, "the level at which an individual learns, understands instructions and solves problems" (Wonderlic,1992, p. 5). Included are items sampled from verbal, quantitative, and analytic domains. Although a time limit is imposed, the Wonderlic behaves more like a power test than a speeded test [7] because the items are presented in ascending order (McKelvie, 1994).

The Wonderlic is very popular in applied settings because of its ease of administration and comprehensive norms combined with ample reliability and validity evidence. Expert reviews have been highly favorable (see Aiken, 1996; Hunter, 1989; Schmidt, 1985; Schoenfeldt, 1985).

The Wonderlic shows test-retest reliabilities ranging from .82 to.94 (Dodrill, 1983; Wonderlic, 1992), and alternate-form reliabilities ranging from .73 to .95 (Wonderlic, 1992). These findings are based on adult working populations, however. Because of restriction of range of ability, college samples should yield lower standard deviations, and therefore lower reliabilities. McKelvie (1989) reported a high internal consistency of .87 (odd-even split-half correlation) in a college sample. The fact that reliability is not increased by relaxing the time requirement (McKelvie, 1994) indicates that the time limit does not inflate the estimate.

In support of concurrent validity, the Wonderlic shows correlations above .80 with longer IQ tests such as the WAIS-R (Dodrill, 1981; Wonderlic, 1992). In fact, Dodrill (1981, p. 668) reported that the Wonderlic IQ scores were within 10 points of the WAIS Full Scale IQ scores in 90 percent of the cases. Of particular note for this report is the fact that correlations are high with measures of both verbal and quantitative abilities (Wonderlic, 1992). Previous studies in college populations have also shown useful predictive validity for college grades (McKelvie, 1994), performance tests (Kennedy, Baltzley, Turnage, & Jones, 1989), and supervisory rankings (Wonderlic, 1992).

RESULTS

Descriptive Statistics

Means, standard deviations, ranges, and reliability coefficient alphas are presented in Table 2.

The values of these statistics in the two samples are virtually identical. Alpha

---

[7] A true speeded test comprises all easy questions.

79

Table 2
Descriptive Statistics

	# Items	Rating Scale	Item Mean	Standard Deviation[a]	Range[a]	Alpha
Global Intelligence Ratings						
Composite Scale[b]	4	5-point	3.66	.67	3.75	.68
			3.24	.69	3.50	.65
Single-item[c]	1	5-point	4.02	.77	4.00	.46[d]
			3.94	.82	4.00	.39[d]
Gough Intellectual	52	True-False	.68	.11	.62.	.72
efficiency (Ie) scale			.68	.10	.54	.67
Trapnell Smart scale	4	9-point	5.45	1.54	8.00	.86
			5.09	1.65	8.00	.88
Hogan Intellect composite	22	True-False	.59	.19	.91	.73
			.58	.20	.91	.77
Intellectance	15	True-False	.62	.21	1.00	.72
			.62	.23	1.00	.76
School Success	7	True-False	.52	.27	1.00	.61
			.48	.26	1.00	.55
Sternberg Behavior	41	9-point	6.24	.82	4.63	.93
Check List (BCL)			6.40	.75	4.10	.92
Verbal Ability	13	9-point	6.02	.96	5.15	.80
			6.12	1.03	5.38	.82
Practical Problem Solving	15	9-point	6.26	.92	5.87	.89
			6.42	.80	4.13	.88
Social Competence	13	9-point	6.45	.84	4.38	.77
			6.67	.72	4.31	.71

Note. Top row of each cell is from Sample 1 ($N = 310$); bottom row is from Sample 2 ($N = 326$).

[a] Standard deviations and ranges are calculated across subject means rather than across item means.

[b] Includes "intelligent" and three conceptually similar items.

[c] "intelligent" (Sample 1); "clear-thinking, intelligent" (Sample 2).

[d] Mean intercorrelations among the four global items.

values for the full scales and subscales are generally quite acceptable, ranging from .61 to .93 in Sample 1 and from .55 to .92 in Sample 2. The reliability of the single-item "intelligent"[8] was estimated from the mean intercorrelation of the four global items from the direct composite.

---

[8] The item was "intelligent, clear-thinking" in Study 2.

Not in the table are the statistics for the Wonderlic IQ test. Our Sample 1 and Sample 2 means (25.5, 26.3) were only slightly higher than the manual norms for college students (Wonderlic, 1992, p. 38): Our SDs (4.41, 4.72), however, were substantially lower than the manual norms of 5.73 for college students.[9] For comparison, note that the norms computed on a representative adult working population (p. 38) exhibited a substantially lower overall mean (21.6) and higher SD (7.1).

Although parallel forms is preferable for estimating the reliability of the Wonderlic, we did not have that information. Instead, reliabilities for the Wonderlic were estimated in two ways. First, we estimated internal consistency directly in our sample with the odd-even split half-reliability used by McKelvie (1989). Our values were .79 and .83 in Samples 1 and 2, respectively. A second calculation involved extrapolating from the appropriate reliability estimates (.90) taken on the broad norm sample (Wonderlic, 1992). Applying the correction formula from Gulliksen (1967, p. 124), to the reduction in standard deviation from 7.12 to 4.41 and 4.72, the alphas in our sample were estimated to be .74 and .77. Using either estimation formula, the reliabilities in our college sample were noticeably lower than in the general population, but certainly within the useful range for research instruments. We can expect that our validities, in turn, will be correspondingly lower than those calculated on the general population.

Range of Responses

As noted earlier, the strong tendency for respondents to claim high levels of intelligence tends to restrict the range of responses, skew the response distribution, and constrain correlations with other variables (McCrae, 1990; Thorndike, 1982). Note that the SDs shown in Table 2 were calculated on the subject means, rather than calculating the means of the item SDs. Given that the latter figures are more relevant to whether or not our subjects were using the entire range of our rating scales, we proceeded to calculate those figures.

Recall that we measured the direct items on 1-to-5 rating scales: The exact distribution of responses was (0, .05, .25, .40, .30) across the two samples. The SD for the single direct item was only .80 and .85 in Samples 1 and 2, respectively. For the four items of the composite direct scale, the average SDs were still small: .93 (Sample 1) and .98 (Sample 2). Compare these values with the average SD of 1.12 for a set of personality items in the same test battery. In short, the direct items did show some restriction in range.

For the indirect measures, the means of the item standard deviations and ranges were not relevant for the two True-False scales (Gough's Ie and Hogan's Intellect composite) and were therefore not calculated. The other two indirect measures were administered in identical 9-point response format, but the variation of the Smart scale items was noticeably greater than the BCL items. The average standard deviations were 1.82 and 1.92 for Trapnell's Smart scale, and 1.59 and 1.53 for the BCL. The average range of the items of the Smart scale was fully 8.00 in both samples, higher than that for the BCL, 7.63 and 7.17.

---

[9] Estimated from the manual norms for mean and women weighted according to gender ratio in our samples.

Intercorrelations Among Predictors

The matrix of intercorrelations among the indirect scales and subscales is presented in Table 3. Note that the four indirect measures (not including subscales) intercorrelate positively but only modestly, with correlation coefficients ranging from .08 to .56 (Sample 1) and from .24 to .47 (Sample 2). Sternberg's subscales intercorrelated quite strongly, with correlation coefficients ranging from .67 to .77 (Sample 1) and from .65 to .68 (Sample 2), while Hogan's subscales intercorrelate modestly, with correlation coefficients of .25 (Sample 1) and .35 (Sample 2).

Table 3

Intercorrelations Among Indirect Scales and Subscales

Scale	1	2	3	4	5	6	7	8	9
1. Gough Ie	—	.08	.47[a]	.42	.32[a]	.43	.41	.36	.39
2. Trapnell Smart	.24	—	.23	.18	.21	.48	.47	.49	.33
3. Hogan Intellect	.27[a]	.29	—	.90	.65	.56	.58	.51	.43
4. Intellectance	.27	.24	.92	—	.25	.51	.50	.49	.38
5. School Success	.24[a]	.25	.68	.35	—	.36	.42	.27	.29
6. Sternberg BCL	.32	.45	.47	.44	.31	—	.88	.93	.89
7. Verbal Ability	.34	.43	.48	.42	.38	.89	—	.72	.67
8. Problem Solving	.29	.50	.43	.44	.22	.89	.67	—	.77
9. Social Competence	.19	.23	.30	.28	.20	.86	.65	.68	—

Note: Correlation coefficients in the upper right of the matrix are from Sample 1 (*N* = 310); coefficients in the lower left are from Sample 2 (*N* = 326). All correlations above .20 are significant, $p < .001$, two-tailed.

[a] Two overlapping items from the Intellectual efficiency and Intellect-School Success scales were assigned to the latter scale for these calculations.

Predictive Validity

Table 4 contains the validities, that is, the correlations of all self-report intelligence measures with Wonderlic test scores. Our baseline validity is that of the single self-rated intelligence item: These values were .20 (Sample 1) and .23

(Sample 2). The corresponding validities for the composite direct measure were slightly higher: .24 (Sample 1) and .26 (Sample 2).

The ability of the four indirect measures to predict IQ test scores was examined in two ways. First we calculated and compared the validities of each predictor; then we performed regression analyses to determine which of the predictors made independent contributions.

Correlations. Table 4 indicates that all four indirect measures achieved significant validities in both samples. Of the indirect measures, Gough's Ie scale performed best, with validities of .20 (Sample 1) and .34 (Sample 2), followed by Trapnell's Smart scale, with validities of .24 (Sample 1) and .25 (Sample 2). Although not as successful overall, the other measures each offered a successful subscale: Hogan's School Success performed well at .19 (Sample 1) and .27 (Sample 2), and so did Sternberg's Verbal Ability subscale, at .24. (Sample 1) and .18 (Sample 2).

Recall that Sternberg (1988) found improved validity via a prototype-weighting procedure (see our introduction). We followed this procedure by having five expert judges (research colleagues) rate each BCL item for diagnosticity of an ideally intelligent person. With strict adherence to Sternberg's method, however,

Table 4
Correlations of Self-Report Measures with IQ Test Scores

Scale	Numbers of Items	Sample 1	Sample 2
Direct Measures			
Single-item[a]	1	.20***	.23***
Composite scale[b]	4	.24***	.26***
Indirect Measures			
Gough Ie	52	.20***	.34***
Trapnell Smart	4	.24***	.25***
Hogan Intellect	22	.15*	.22***
Intellectance	15	.08	.13*
School Success	7	.19**	.27***
Sternberg BCL	41	.20***	.13*
Verbal Ability	13	.24***	.18**
Practical Problem Solving	15	.17**	.10
Social Competence	13	.14*	.04

Note: *p < .05; **p < .01, ***p ≤ .001, two-tailed. Sample size ranges from 274 to 301 (Sample 1) and from 241 to 265 (Sample 2) due to the subject matching across the three sources of data (i.e., direct, indirect, and IQ measures).

[a] "Intelligent" (Sample 1); "clear-thinking, intelligent" (Sample 2).

[b] Refers to the direct, global intelligence ratings including "intelligent" and three conceptually similar items.

we found no validity improvement. When we simply weighted without standardizing the BCL item scores, the correlations did improve slightly from .20 to .23 (Sample 1) and from .13 to .17 (Sample 2).

A follow-up set of regression analyses was conducted to determine the predictive power of the Intellect and BCL subscales. When the three Sternberg subscales alone were simultaneously force-entered, they accounted for a total of 6 and 4 percent of the variance in our two samples. A similar forced-entry with the Intellect subscales accounted for a total of 7 percent of the variance in both samples. As might be expected, regression on the subscales accounted for more of the variance than that achieved by the composite Intellect or BCL scores.

A Two-Factor Organization of Strategies

Table 5 summarizes the key data for this report by displaying the mean validities of the four categories of measures of self-report intelligence. The performance of direct measures can easily be compared with those of indirect measures for both single items and the aggregated scales.

Table 5

Correlations of Four Types of Predictors with IQ Test Scores

	Aggregation Strategy			
	Single Item		Aggregated Items	
Directness Srategy				
	Sample 1	Sample 2	Sample 1	Sample 2
Direct	.20	.23	.24	.26
Indirect	.07	.07	.18	.24

Note: Single-item validities are correlations of IQ scores with the single item, "intelligent" (Study 1) or "clear-thinking, intelligent" (Study 2); the single-item indirect validities are mean item validities across all 119 items of the four indirect measures. The aggregated items direct validities are based on the four-item direct composite measure. The aggregated items indirect validities are the mean full-scale validities across all four indirect measures.

The entries in Table 5 (across the rows) are as follows: The single-item direct validities are the correlations of IQ with the single item "intelligent" (Study 1) or "clear-thinking, intelligent" (Study 2). The aggregated direct entries are the validities of the four-item direct scales including "intelligent" and closely related items. The single-item indirect validities are mean item validities across all 119 items of the four indirect scales. Finally, the aggregated indirect validities are the mean full-scale validities across all four indirect measures.

In the case of the direct measures, aggregation boosts the validities from .20

BEST COPY AVAILABLE

and .23 to .24 and .26, respectively. This small improvement with aggregation was disappointing. Validity prophecy formulas,[10] for instance, would predict values of .26 and .30 for a measure comprising four items equivalent to our baseline single items. Apparently, the validities of our additional items did not parallel those of the original item (intelligent). Despite our best efforts to select conceptually similar items, aggregation provided only modest improvements in the validity of direct measures.

For the indirect scales, however, Table 5 reveals a dramatic effect for aggregation. Here the comparison is between the mean of the 119 item validities (.07 and .07) and the mean of the four full scale validities. Broken down scale by scale, aggregation raised the validities from .05 to .20, .20 to .24, .05 to .08, and .11 to .20 for Ie, Smart, Intellect, and the BCL, respectively, in Sample 1, and from .07 to .34, .21 to .25, .09 to .13, and .06 to .13 in Sample 2. In short, all indirect scales benefited from aggregation.

A Closer Examination of the Indirect Measures Using an Empirical Approach

Table 6 presents the 10 best performing items from the indirect scales, as defined by consistently good validities. Every measure is represented in the top 10. The fact that the Ie scale contributes the largest number of representatives probably derives from the fact that it has the most items and therefore the greatest opportunity to capitalize on chance. It is noteworthy that the items with the highest validities are those related directly to mental ability: After all, the rationale behind the creation of these indirect measures was that indirect items should exceed the validity of the more blunt, direct items such as "intelligent" or "smart."

All four Smart items performed well in both samples, and two made the top 10. Note that the top Intellect items are from the School Success subscale and that all concern reading. Also note that the top items of the BCL both come from the Verbal Ability subscale.

Summarizing the contents of Table 6, then, it appears that the top-performing items in the indirect scales were either (1) direct ability-related items, (2) indirect items about ability (i.e., Smart scale), or (3) items about reading behavior. If these 10 items are combined into a new "best items" composite scale, the correlation with IQ test is .34 in Sample 1 and .38 in Sample 2.[11]

---

[10] The prediction formula when only one measure is lengthened is presented by Thorndike (1982, p. 153). An infinite number of equally good items would remove all unreliability to yield upper limits of .30 and .37 in the two samples.

[11] These values are likely to be overestimates because of capitalization on chance. Unfortunately, cross-validation from one sample to the other is not feasible because the items were chosen on the basis of consistent performance across both samples.

Table 6
The Top 10 Item Validities from Indirect Measures

Sample 1	Sample 2	Scale	Item Content
.27	.32	BCL	reads with high comprehension
.28	.26	BCL	has a good vocabulary
.27	.20	Intellect	as a child I was always reading
.21	.21	Ie/Intellect	I am quite a fast reader
.18	.23	Ie/Intellect	I read at least ten books a year
.20	.25	Smart	Is considered a very "brainy," scholarly person
.21	.21	Ie	I was a slow learner in school
.20	.25	Smart	considered exceptionally or unusually intelligent
.20	.20	Ie	I seem to be at least as capable and smart as most others around me
.22	.17	Intellect	I would rather read than watch TV

Note: Values > .15 are significant at .01, while those > .20 are significant at .001, two-tailed. *N*= (275, 265). "Ie" refers to Gough's Intellectual efficiency scale; "Smart" refers to Trapnell's scale; "Intellect" refers to Hogan's scale; "BCL" refers to Sternberg's Behavior Check List.

A Closer Examination of the Indirect Measures Using a Theoretical Approach

Having discovered that the best performing items of the Sternberg, Hogan, and Gough scales were, in fact, the more direct and ability-related items, we decided to categorize the items of these indirect measures theoretically. Four a priori categories were considered: mental abilities, personality-related, behaviors, and interests. Two judges showed 95 percent agreement on classification.

For each category, the mean validities were calculated; they are presented here in Table 7. Mean item validities were .14 (Sample 1) and .12 (Sample 2) for the ability items. Means for next three categories (personality, interests, and behaviors) were in the .03 to .06 range — all substantially lower than the ability-related items.

Best of all, however, was the set of items addressing reading habits: "I read at least 10 books a year" from the Ie and Intellect scales, "As a child I was always reading" from Hogan's Intellect composite scale, and "reads widely" and "Sets aside time for reading" from the BCL. In fact, these reading items showed exceptional mean validities of .19 (Sample 1) and .18 (Sample 2).

In sum, it appears that items directly related to mental ability and items about reading habits outperformed the other item content categories in predicting IQ test scores. The other categories of items show positive, but low, item validities.

Table 7
Validities of Indirect Items: Means Within Content Category

Category of Items	# Items[a]	Sample 1	Sample 2
Ability-related	23	.14	.12
Personality-related	51	.05	.06
Interest-related	26	.03	.06
Behavior-related (Non-reading)	8	.05	.05
Reading	5	.19	.18

Note: *N*s = 275 (Sample 1) and 265 (Sample 2). Each entry is the mean of all item validities for each category.

[a] Number of items refers to the number of item validities used in calculating the mean item validity.

Nonetheless, they can be aggregated, as in the case of the Ie scale, to reach a reasonable level of validity.

DISCUSSION

We set out to evaluate whether IQ can be measured by proxy. That is, can the handy self-report format be used as a substitute for a cumbersome IQ test? Because the validity of a single self-rating of intelligence has not proved adequate, researchers have advocated a number of strategies for improving validity, namely, indirectness, aggregation, and prototype-weighting. Our results indicated that aggregated, direct measures were the most effective, but none could consistently exceed .30. Prototype weighting had minimal impact.

Performance of Direct Measures

We began by establishing the validity of our baseline, that is, a single face-valid self-rating of intelligence. In two large samples, the single item showed validities of .20 and .23 — values that are typical of previous studies. Some studies have reported higher validities but most of those were high-school or other samples with a wide range of talent. Competitive college samples, such as our own, suffer from a restricted range of ability, which limits potential validity values. In any case, our modest baseline values left plenty of room for improvement via aggregation.

The empirical benefits of aggregation were evaluated by pooling the single item with other intelligence-related items[12] in a composite direct measure. An improvement in validity was observed with the addition of the 3-4 items most synonymous/antonymous to “intelligent,” namely, smart, clever, simple, and not gifted. The lack of improvement beyond four items suggests that further aggregation added more noise than valid variance. Our battery of items contained 13 items selected for relevance to mental ability. But few of these were able to capture that facet of self-perception linked to IQ.

Even the validities of the 4-item composites (.24, .26) do not match the values predicted by the validity prophecy formula (26, .31) based on projecting the validity of the single item "intelligent" (Thorndike,1982, p. 152). Clearly that item plays a unique role, both conceptually and empirically.

Performance of Indirect Measures

Indirect measures promised to surpass the performance of direct measures by providing a less threatening, less evaluative assessment atmosphere. In terms of predicting IQ test scores, that promise was not fulfilled in our data. Given that the results for each indirect measure raised different issues, however, we will consider them one by one.

Gough Intellectual efficiency (Ie) scale. Recall that the Ie scale was constructed in a contrasted-groups fashion by selecting CPI items that correlated with an IQ test (Gough, 1953). Given that it was developed decades ago on California high-school students, its success in our contemporary Canadian college sample — that is, validities of .20 and .34 — might be considered remarkable. Nonetheless, a full 52 items were required to achieve those full-scale validities because the mean item validity was low. Of course, True-False items are expected to show lower validities (but faster administration times) than corresponding Likert items.

Contrary to the original intent of the Ie, however, it was primarily the direct mental-ability items that correlated with IQ. The remaining items, concerned primarily with confidence and adjustment, were not as successful, although all items were originally selected because they correlated with IQ tests. Why the confidence and adjustment item validities did not replicate is difficult to say. It is understandable that distractability related to maladjustment could hamper performance on IQ tests independent of actual ability. And this handicap of maladjustment may be more true in high-school samples (where item-selection took place) than in the college samples where we chose to validate the items.

Perhaps the four-decade gap in culture is somehow responsible. Even across a 10-year time span, Paulhus and Landolt (1994) found that the criteria for nominating intelligent people had changed noticeably whereas the criteria for the concept of intelligence had not changed. We suspect that, when criterion groups rather than rational methods are used to develop scales, items measuring temporary societal influences are more likely to intrude.

Sternberg's Behavior Check List. As a unit, the BCL showed only modest

---

[12] The items were selected by their conceptual similarity to "intelligent" (gifted, smart, clever, etc.).

predictive efficacy — slightly higher if an item-weighting procedure was applied. One of the subscales, the Verbal Ability subscale, was effective. Sternberg et al. (1981) found the same pattern. Our detailed item analyses revealed that the high-validity items carrying the subscale were those concerning mental ability and items about reading habits.

Although the ability of the BCL to predict IQ was not impressive for a 41 Likert-item measure, we must call attention to its original purpose. Sternberg intended the BCL not as a *proxy* for IQ tests, but as a *supplementary measure*: It was designed to be administered along with an IQ test to tap components of intelligence that IQ tests were not capable of measuring (Sternberg, 1988, p. 239). From this perspective, high correlations with IQ tests should not be expected.

This supplementary role of the BCL is consistent with Sternberg's long-standing complaint that IQ tests measure only a limited part of lay conceptions of intelligence. Recently, we have followed up this notion in our work on "non-test intelligence" (Lysy & Paulhus, 1996). By partialing IQ and self-presentation out of self- and peer-ratings of intelligence, we formed a self-residual and a peer-residual to represent that part of intelligence that is "beyond IQ." We then correlated the residuals with a battery of personality and interest measures. The top correlates of the self-residual were self-rated conscientiousness and openness, self-esteem, the Intellectual efficiency scale, and the Smart scale, while the top correlates of the peer residual were peer-rated conscientiousness, openness, physical attractiveness, and athletic ability. The different correlates of self and peer suggest that "non-test intelligence" is largely a perceiver-dependent idiosyncrasy. There was, however, a small overlapping component indicating that self and others systematically misattribute intelligence to those who are conscientious and open. This component may be that facet of "true intelligence" that is not represented in IQ tests.

Hogan's Intellect composite. As a whole, the Intellect composite showed only a modest ability to predict IQ scores. Obscured in this overall figure, however, is the fact that the two subscales showed dramatically different validities. Recall that the Intellectance subscale was designed to capture an unconventional, creative conception of intellect whereas School Success was aimed at the more conventional goal-oriented conception of intellect. Our results support this distinction in that School Success was a distinctly better predictor of IQ with validities of .19 and .27 in our two samples. In fact, these are underestimates because we used only a seven-item version. The validity of a 21-item version, as predicted by the prophecy formula (Thorndike, 1982), would have been .22 and .32 in our two samples.

Trapnell's Smart scale. The newest instrument in the study, Trapnell's (1994) Smart scale, performed well. It was designed to reduce range-restriction in two ways: (1) by diminishing the desirability of claiming the item and (2) by shifting the implied locus of evaluation from self to others. As intended, the Smart scale did show a reduced range restriction: Subjects utilized almost the entire range of the 9-point scale — noticeably more than the range of Sternberg's BCL.

The Smart scale is certainly efficient, requiring only four items to match or even outperform the other indirect measures. It is now evident, however, that the success of the Smart scale did not derive from its indirect nature. Direct composites with four items of similar content (smart, clever, etc.) worked just as well as the Smart scale. Therefore its success was more likely a function of content rather than of Trapnell's strategic contextualizing of the items.

The Content of Predictors

This scale-by-scale analysis of successful items helped clarify the source of their success. Although the four inventories derived from four dramatically different domains, the successful items within each were almost entirely ability-related. Of course, the direct measures were designed to address ability directly. But in the case of indirect measures, it is certainly ironic that their most direct items work best. This finding has the added benefit of refuting a potential alternative explanation for our finding of lower validities for indirect measures than direct measures, namely, that the indirect measures were administered later in a separate test battery from the direct measures. But even those direct items included in the same battery as the indirect still outperformed the indirect items.

Out of all remaining item-content areas, the only one yielding consistently high validities was an interest in reading.[13] Why a lifelong enjoyment of reading is associated with achieving high scores on IQ tests is not clear. Many educational psychologists argue that reading behavior permanently boosts mental abilities (Rayner & Pollatsek, 1989) and is rightfully encouraged. Of course, other causal sequences are possible. High intelligence might make reading more enjoyable (Hogan & Hogan, 1992). Or third variables such as social class or openness to experience might nurture both (McCrae & Costa, 1985).

The Value of Aggregation and Weighting

Administration of the single item "intelligent" is certainly efficient given the practical costs of adding more items to a test battery. And, across all items administered in our studies, it was the most consistently valid. Addition of other direct items improved validities only up to five items. Beyond that, returns were marginal.[14] Apparently, the items linked to IQ scores have a limited semantic scope. The fact that the Ie scale benefited most from aggregation suggests that this strategy aids true-false more than Likert-item composites. It is understandable that dichotomous items, though potentially as valid as Likert items, require more aggregation because of lower item reliability.

---

[13] Interestingly, an instrument recently developed to predict school success contains the same two categories of predictors (Giddan, Jurs, Andberg, & Bunnell, 1996).

Sternberg et al. (1981) reported substantial improvement in BCL validities via a correlation-with-prototype approach. As we showed in the Introduction, this approach is simply a form of weighting procedure, that is, counting certain items more than others in calculating total scale scores.[15] The traditional psychometric wisdom is that unit weighting of items selected from regression or factor analysis is preferable to any other weighting: That wisdom was not refuted by our data.

A general-purpose inventory, such as Sternberg's BCL, however, may represent an interesting exception to that wisdom. Here, an eclectic set of items is to be used for a variety of purposes, in this case, all germane to intelligence (see Cornelius et al., 1989). For predicting IQ, our analyses showed that the Social Competence items should be given zero weights; in predicting some other criterion, a different set of items might be zero-weighted. In a sense, the weightings are used to "unload" the items that are irrelevant for the current purpose. Thus a heterogeneous set of items can be retained but weighted in different ways to predict different criteria.

According to this argument, the only instruments in our package with potential for improvement via weighting are the BCL and the Ie scale. Unfortunately, although we tried various weighting procedures, we achieved only minimal improvements. Nonetheless, in appropriate instruments, such weighting could prove useful.

Putting Our Results in Perspective

Are self-report measures useful as proxies for IQ tests in college samples? Our data suggest not. Given that the validity of an ideal proxy measure would be upwards of .55[16] in college samples, our validity cap of .30 is disappointing. We tried out the best available measures, as well as the most highly touted improvement strategies.

Limitations in the criterion? The criterion measure, the Wonderlic IQ test, does not appear to be at fault. Previous studies have shown sufficient construct validation in college populations (e.g., Kennedy et al., 1989; McKelvie, 1994; Paulhus & Morgan, 1997; Wonderlic, 1992). Rather than being inappropriate for measuring IQ in college samples, its lackluster performance here is directly

---

[14] It seems that aggregation may pay off less in assessing intelligence than in assessing personality. However, our selection of intelligence items was not systematic enough to make such a strong claim here.

[15] If the negatively keyed items have not yet been reversed, then the primary effect of weighting is simply to reverse these items.

[16] This estimate begins with the median correlation (.83) of the Wonderlic with other IQ tests in general populations (Wonderlic, 1992). Instead of 7.12, the general population SD, the mean SD of our two samples was only 4.6. Adjustment of the validity for this restriction in range (Cohen & Cohen, 1983) yields .55.

attributable to its low standard deviation. It performed no better and no worse than any standard IQ test would have in this situation.

Contamination of self-reports? So why the poor correspondence between self-rated intelligence and IQ tests? As Sternberg (1998) has noted, correspondence is limited by the common tendency to base one's self-perceived intelligence on abilities different from those tapped by IQ tests. In addition, discordance is to be expected because of motivated as well as unmotivated ignorance (Paulhus, 1986). The motivated portion involves inflated self-perceptions due to narcissism or self-deception. Previous research shows that this component contributes even more than IQ scores — perhaps 20 percent of the reliable variance in self-perceived intelligence (Gabriel, Critelli, & Ee, 1994; Paulhus, Yik, & Lysy, 1996). This motivated component also includes idiosyncratic definitions of intelligence designed defensively to match the raters' own abilities and therefore ensure that they are intelligent (Dunning & Cohen, 1992). The unmotivated portion of ignorance may include a lack of interest, concern, or insight into such matters (Campbell & Lavallee, 1993).

Restriction of range? Finally, we must remind readers of the severe handicap placed on all the validities reported here. The restriction of range created by our use of college samples is likely to have diminished all validities as a function of the reduced variances (see Cohen & Cohen, 1983). Compared with the SD of 4.6 that we found for our IQ test, SDs of 7.1 are more typical of the general populations (Wonderlic, 1992). Adjusted for restriction of range, our baseline validities for the single item "intelligent" (.20-.23) would have reached .30-.35. Similarly, instead of our ceiling of .30 for aggregated instruments, we could have achieved values of .40-.45 in the normal population. The latter values appear strong enough to be useful in research, if not in diagnosing individuals.

Limitations of IQ tests? Self-reports of intelligence, we argue, should not be evaluated solely in terms of potential as proxies for IQ tests. Given that lay perceivers typically hold that there is more to intelligence than IQ, our participants may well have based their self-ratings on their creativity, their interpersonal sensitivity, their musical ability, or their self-insight — none of which are tapped by the Wonderlic. And we agree with the view of expert commentators such as Sternberg and Gardner that we must tie scientific conceptions of intelligence more closely to such lay conceptions.

Such arguments suggest an alternative criterion for evaluating self-reports of intelligence: the perceptions of knowledgeable peers. In support of this argument, we have elsewhere reported evidence that self-ratings predict peer-ratings of intelligence independent of IQ scores (Lysy & Paulhus, 1996). That is, some portion of observers' perceptions of intelligence is detectable by self and observers but not by IQ tests. Thus self-report measures of intelligence have validity beyond their use as proxy IQ measures. From this perspective, it would actually be surprising to find high correlations between IQ tests and perceptions of intelligence.

Some Promising Avenues

We see several potential avenues for clarifying the links between test performance and self-perceptions of intelligence. First is the development of new intelligence tests to encompass more of everyday conceptions of "intelligence." To the extent that test content corresponds to everyday conceptions, then associations should be higher. Wagner and Sternberg (1986) have pursued this avenue by developing objective measures of practical intelligence. Salovey and Mayer's (1990) "emotional intelligence" is another measure that shifts the conceptual borders of intelligence toward everyday conceptions.

A second avenue for future research is clarifying and perhaps improving the other side of the relationship, namely, the self-perceptions. What cues are people using to judge their intelligence? The lens model is proving profitable in specifying proximal cues, that is, objective behaviors that trigger attributions of intelligence (Reynolds & Gifford, 1996). We too are examining matches and mismatches between self- and peer-perceptions of intelligence and their correlates (Lysy & Paulhus, 1996). This research should help specify the missing content in current self-report measures.

In a third avenue of research, we have attempted to deal with the self-presentation typical of self-reported intelligence. The Overclaiming Questionnaire exploits a sophisticated methodology with great potential as a proxy IQ test (Paulhus, Bruce, & Lysy,1996). Respondents are asked to rate their familiarity with a wide range of people, places, books, events, and so on. Because 20 percent of the items are fictitious, signal detection statistics can be used to separate accuracy from bias. In a series of college samples, the signal detection accuracy parameter (*d'*) correlated .44-.50 with scores on an IQ test. Considering that these were college samples, the validities are quite promising.

Finally, we encourage further research on the indirect measures studied here. Their greatest potential asset has never been directly tested: they may actually outperform direct measures in ego-threatening administration conditions. Another issue worthy of study is whether the direct ability items work only when interspersed with a variety of other items.

CONCLUSIONS

The present paper constitutes the most comprehensive examination of self-reports of intelligence to date. We have organized the available measures into four categories of self-rated intelligence to investigate the effects of employing indirect versus direct measures, and the effects of aggregation, on predicting objectively scored intelligence. Administration of these measures to two large samples led us to a few key conclusions.

1. Both direct and indirect self-report measures of intelligence can reliably predict IQ scores. Because of the restricted range of abilities in competitive col-

lege samples, however, the validity limit appears to be .30.

2. Direct items about global mental abilities are more valid than indirect items. The one clear exception is the high validities of indirect items referring to enjoyment/frequency of reading.
3. Aggregation of global ability items is beneficial up to a point. With the exception of reading items, aggregation doesn't appear to help beyond 4-5 core items referring directly to close synonyms/ antonyms of intelligence (e.g., smart, clever, simple, not gifted).
4. Prototype weighting is helpful only for excluding ineffective items in an inventory.
5. Among available measures, the most effective predictors of IQ scores were Gough's Intellectual efficiency and Trapnell's Smart scale. Equally effective were Hogan's School Success scale and Sternberg's Verbal Ability scale.

As a whole, our verdict is pessimistic about the utility of self-reports as proxy measures of IQ in college samples. Our verdict is more optimistic about their utility for assessing intelligence as a broader concept, particularly in the general population. Either way, researchers who require some proxy IQ test for their research should benefit from the guidelines we have provided here.

Correspondence should be addressed to Del Paulhus, Department of Psychology, University of British Columbia, Vancouver, Canada, V6T 1Z4. Email may be sent to dpaulhus@cortex.psych.ubc.ca.

References

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121,* 219-245.

Aiken, L. R. (1996). *Assessment of intellectual functioning* (2d ed.). New York: Plenum Press.

Bailey, R. C., & Hatch, V. (1979). Interpersonal perceptions of intelligence in late childhood and early adolescent friendships. *Journal of Genetic Psychology, 135,* 109-114.

Bailey, R. C., & Mettetal, G. W (1977). Perceived intelligence in married partners. *Social Behavior and Personality, 5,* 137-141.

Block, J. (1961). *The Q-sort method in personality assessment and psychiatric research.* Springfield, IL: Charles C. Thomas.

Block, J. (1971). *Lives through rime.* Berkeley, CA: Bancroft Books.

Borkenau, P (1993). To predict some of the people more of the time: Individual traits and the prediction of behavior. In K. H. Craik, R. Hogan, & R N. Wolfe (Eds.), *Fifty years of personality psychology: Perspectives on individual differences* (pp. 237-249). New York: Plenum Press.

Borkenau, P., & Liebler, A. (1993). Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence. *Journal of Personality and Social Psychology, 65,* 546-553.

Campbell, J. D., & Lavallee, L. F. (1993). Who am I? The role of self-concept confusion in understanding the behavior of people with low self-esteem. In R. F Baumeister (Ed.), *Self-esteem. The puzzle of low self-regard* (pp. 3-20). New York: Plenum Books.

Cohen, J., & Cohen, P. (1983). *Applied multivariate regression/correlation for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cornelius, S. W., Kenny, S., & Caspi, A. (1989). Academic and everyday intelligence in adulthood: Conceptions of self and ability tests. In J. D. Sinnott (Ed.), *Everyday problem-solving: Theory and applications* (pp. 191-210). New York: Praeger.

DeNisi, A. S., & Shaw, J. B. (1977). Investigation of the uses of self-reports of abilities. *Journal of Applied Psychology, 62,* 641-644.

Dodrill, C. B. (1983). Long term reliability of the Wonderlic Personnel Test. *Journal of Consulting and Clinical Psychology, 51,* 316-317.

Dunning, D., & Cohen, G. L. (1992). Egocentric definitions of traits and abilities in social judgment. *Journal of Personality and Social Psychology, 63,* 341-355.

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality, 51,* 360-392.

Gabriel, M. T., Critelli, J. W, & Ee, J. S. (1994). Narcissistic illusions in self-evaluations of intelligence and attractiveness. *Journal of Personality, 62,* 144-155.

Giddan, N. S., Jurs, S. G., Andberg, M., & Bunnell, E (1996). Noncognitive long-term prediction of college grades by the Academic Performance Scale. *Assessment, 3,* 91-98.

Gough, H. G. (1953). A nonintellectual intelligence test. *Journal of Consulting Psychol*ogy, 17, 242-246.

Gough, H. G. (1996). *California Psychological Inventory* (3rd. ed.). Palo Alto, CA: Consulting Psychologists.

Gulliksen, H. (1967). *Theory of mental tests.* New York: Wiley.

Hogan, R., & Hogan, J. (1992). *Hogan Personality Inventory manual.* Tulsa, OK: Hogan Assessment Systems.

Hunter, J. E. (1989). *The Wonderlic Personnel Test as a predictor of training success and job performance.* Technical report, Department of Psychology, Michigan State University.

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domain, observability, evaluativeness, and the unique properties of the self. *Journal of Personality, 61,* 521-551.

Kennedy, R. S., Baltzley, D. R., Turnage, J. J., & Jones, M. B. (1989). Factor analysis and predictive validity of microcomputer-based tests. *Perceptual & Motor Skills, 69,* 1059-1074.

Lysy, D. C., & Paulhus, D. L. (1996, August). *Beyond IQ: The search for non-test intelligence.* Paper presented at the meeting of the American Psychological Association, Toronto.

McCrae, R. R. (1990). Traits and trait names: How well is openness represented in natural languages? *European Journal of Personality, 4,* 119-129.

McCrae, R. R., & Costa, P. T. (1985). Updating Norman's "Adequate Taxonomy": intelligence and personality dimensions in natural language and in questionnaires. *Journal of Personality and Social Psychology* 49, 710-721.

McKelvie, S. J. (1989). The Wonderlic Personnel Test: Reliability and validity in an academic setting. *Psychological Reports*, 65, 161-162.

McKelvie, S. J. (1994). Validity and reliability findings for an experimental short form of the Wonderlic Personnel Test in an academic setting. *Psychological Reports*, 75, 907-910.

Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaire* (pp. 143-165). New York: Springer-Verlag.

Paulhus, D. L., Bruce, M. N., & Lysy, D. C. (1996). *The Over-Claiming Questionnaire (OCQ)*. Unpublished instrument, University of British Columbia.

Paulhus, D. L., & Landolt, M. (1994, June). *Differential processes in accessing concepts of intelligence.* Paper presented at meeting of Canadian Psychological Association, Penticton, Canada.

Paulhus, D. L., & Martin, C. L. (1987). The structure of personality capabilities. *Journal of Personality and Social Psychology*, 52, 354-365.

Paulhus, D. L., & Morgan, K. L. (1997). Determinants of perceived intelligence in leaderless groups: The dynamic effects of shyness and familiarity. *Journal of Personality and Social Psychology*, 72, 99-107.

Paulhus. D. L., Yik, M.S.M, & Lysy, D. C. (1996, August). *Self- and peer-ratings of intelligence: Accuracy or self-presentation?* Paper presented at the meeting of the American Psychological Association, Toronto.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading.* Englewood Cliffs, NJ: Prentice-Hall.

Reilly, J. & Mulhern, G. (1995). Gender differences in self-estimated IQ: The need for care in interpreting group data. *Personality and Individual Differences, 18,* 189-192.

Reynolds, D., & Gifford, R. (1996). *Measured and judged intelligence: A Brunswik lens analysis of verbal and nonverbal cues.* Manuscript in progress, University of Victoria.

Salovey, P., & Mayer, J. D. (1989-1990). Emotional intelligence. *Imagination, Cognition, and Personality, 9,* 185-211.

Schmidt, F. L. (1985). Review of the Wonderlic Personnel Test. In J.V. Mitchell (Ed.). *Ninth mental measurements yearbook* (pp. 1755-1757). Lincoln, NE: Buros Institute of Mental Measurement.

Schoenfeldt, L. F. (1985). Review of Wonderlic Personnel Test. In J. V. Mitchell (Ed.), *Ninth mental measurements yearbook* (pp.1757-1758). Lincoln, NE: Buros Institute of Mental Measurement.

Sternberg, R. J. (1988). *The triarchic mind: A new theory of human intelligence.* New York: Penguin Books.

Sternberg, R. J., Conway, B. E., Ketron, J. L., & Bernstein, M. (1981). People's conceptions of intelligence. *Journal of Personality and Social Psychology, 41,* 37-55.

Trapnell, P. D., & Scratchley, L. (1996). *Predictors of intellectual performance.* Unpublished data, University of British Columbia.

Thorndike, R. L. (1982). *Applied psychometrics.* Boston: Houghton-Mifflin.

Trapnell, P. D. (1994). Openness versus intellect: A lexical left turn. *European Journal of Personality, 8,* 273-290.

Wagner, R. K., & Sternberg, R. J. (1986). Tacit knowledge and intelligence in the everyday world. In R. J. Sternberg & R. K. Wagner (Eds.), *Practical intelligence: Nature and origins of competence in the everyday world (pp.* 51-83).New York: Cambridge University Press.

Welsh, G. S. (1975). *Creativity and intelligence: A personality approach.* Chapel Hill, NC: Institute for Research in Social Science.

Wonderlic, E. F. (1992). *Wonderlic Personnel Test and scholastic level exam user's manual.* Libertyville, Il: Wonderlic Personnel Test, Inc.

BEST COPY AVAILABLE

Gifted — Through Whose Cultural Lens?
An Application of the Postpositivistic Mode of Inquiry

Jean Sunde Peterson, Purdue University

Using a postpositivistic method of inquiry, this study examined themes that emerged in the language of Latino, African American, Native American, immigrant Asian, and low-income Anglo individuals as they nominated individuals for a hypothetical gifted program. Reflected values differed from group to group and differed considerably from those reflected in classroom teachers' language in an earlier study. The researcher theorized that value orientations of mainstream teachers, who make referrals for programs after initial screening procedures, preclude their valuing behaviors deemed valuable by nonmainstream individuals and also inhibit behaviors deemed gifted by teachers. Findings suggest that the mode of inquiry can be useful for those who seek new ways to conceptualize giftedness and who seek to explain phenomena, such as the underrepresentation of nonmainstream groups in gifted education.

Introduction

Classroom teachers may be pivotal players in the continuing underrepresentation of nonmainstream children in programs for *gifted students,* according to two studies exploring value orientations of dominant-culture teachers and nonmainstream students. Both studies used a postpositivistic mode of inquiry. The first (Peterson & Margolin, 1997) analyzed the language of teachers regarding conceptualizations of *giftedness* as they nominated students for a gifted program. The second, which is the main focus of this paper, also focused on language, exploring various cultural groups' conceptualizations as they made similar nominations. In the initial study, analysis regarding conceptualizations of giftedness precipitated a focus on cultural values, as reflected in the themes that emerged. In the second study, themes and reflected values differed considerably from group to group and between the nonmainstream groups and the mainstream teachers. The differences argue against a narrow conceptualization of giftedness and suggest that differing value orientations may preclude affirmation of nonmainstream students in the classroom and their nomination for special programs. The studies also demonstrate the usefulness of the postpositivist mode of inquiry, especially in exploring complex and enduring issues, such as definition and equity.

Reprinted from *Journal for the Education of the Gifted*, vol. 22, no. 4, 1999.

In more than 90 percent of states, classroom teachers participate in identification processes *(*Adderholdt-Elliot, Algozzine, Algozzine, & Haney, 1991). Their referrals bear heavily on whether programs eventually include poor and minority students, whose proportionate numbers have been increasing in the United States (McKenney & Bennett, 1994) and who are among those with high ability who most need programs to develop their latent potential (Richert, 1985). Nonmainstream children and adolescents may not be recognized for their abilities during initial screening procedures, which are likely to include cut-off scores on a standardized achievement battery or on an intelligence test (Maker, 1996; *National Excellence,* 1993), even in school districts specifically attempting to identify underserved populations (Callahan, Tomlinson, & Pizzat, 1993). Those tests, to many, define giftedness (Hoge, 1988). In keeping with definitional structures of programs (VanTassel-Baska, 1991), test scores tend to be given disproportionate weight when multiple measures are used and are among measures that may be combined inappropriately. They may, in fact, be used only after disadvantaged students have already been excluded (Richert, 1997). According to Richert (1992), most efforts to use data beyond achievement measures during identification processes are cosmetic; and "the more measures that are used and combined inappropriately the more likely it becomes that disadvantaged students... will be excluded" (p. 7). In addition, parents of minority children are not as likely to request evaluations for their children for possible placement as are mainstream parents (Scott, Perou, Urbano, Hogan, & Gold, 1992). Prospects for participation in programs may, therefore, rest in the hands of classroom teachers when they are asked to make referrals.

If teachers are from the dominant culture, no matter how well intentioned they are, their cultural value orientation may interfere with referring nonmainstream children. When classroom teachers are provided checklists for marking *gifted behaviors* (eg., Kingore, 1990; Renzulli & Hartman, 1971), listed behaviors may reflect dominant-culture values, may include behaviors that are actually discouraged in nonmainstream cultures, and may lack reference to characteristics that give a cultural group its identity (Montgomery, 1989). Programs that identify and serve students on the basis of strengths in specific domains (cf. Renzulli & Delcourt, 1986) may also reflect dominant-cultural values, which affect which domains receive focus and the assessment of how talent is expressed and assessed within those domains. Richert (1987) noted that "only those students who have developed certain modes of thinking rewarded in schools will be nominated" (p. 152).

Values Literature in Gifted Education

Attention to cultural values in gifted education is not new. More than two decades ago, scholars pointed out the inadequacy of psychometric measures to identify nonmainstream groups equitably (Baldwin, 1978; Bernal, 1977; Bruch

& Curry, 1978; DeAvila, 1974; Torrance, 1977; Witty, 1978) and promoted community-based definitions and assessments (Bernal, 1974; Passow, 1972). Since then, they and others have continued to draw attention to inappropriate selection processes and narrow conceptualizations (e.g., Bernal, 1981; Feldhusen, 1998; Ford, 1996; Passow & Frasier, 1996; Maker, 1996; Richert, 1987).

Gardner's (1983) conceptual framework for viewing intelligence acknowledges the importance of cultural context in regard to particular competencies. Mistry and Rogoff (1985) saw a relationship between cultural belief systems and valued skills, and Passow and Frasier (1996) emphasized that

> Talents of minority and economically disadvantaged students are not of a different order nor of a lower standard. By its very nature, any culture or subculture can encourage or inhibit an individual's behavior by the rewards or sanctions it provides. (p. 199)

Scholars have also focused on values of specific groups in terms of culturally relevant definitions, identification, or programs. Some representative cultures and authors are the following: American Indian (Tonemah, 1991); Hawaiian (Buchanan, Javier, Sing, & Plunkett, 1993); Maori (Reid, 1992); Asian American (Shen & Mo, 1990); African American (Ford, 1996); Mexican American (Duran & Weffer, 1992); and rural (Kleinsasser, 1986).

Attention to values is often focused on students already identified for programs (e.g., Colangelo & Parker, 1981; Grant, 1995; Howard-Hamilton & Franks, 1995; Lubinski, Schmidt, & Benbow, 1996). Buckley (1992) concluded that schools selected students whose families concurred with their conceptions of giftedness: The central categories of IQ, creativity, and motivation emerged when parents of identified children were asked for attributes of giftedness to be considered in programs for the gifted. These studies have generally not addressed these values as cultural. Discussions of new paradigms notwithstanding (e.g., Feldman, 1992; Maker, 1996; Treffinger & Feldhusen, 1996), the gifts being sought remain largely performance based and related to school achievement, emphases that may be complexly problematic in light of differing cultural values. Scrutiny of traditional emphases can provoke questions "for which there are no reasonable answers within current ways of thinking about giftedness" (Feldman, p. 90).

In regard to cultural values and identification procedures, Torrance (1978) wrote, "If educators are really interested in identifying gifted and talented students in minority groups they will direct their searches to those characteristics that are valued by the particular minority groups" (p. 30), while also noting that most schools do not provide opportunity for displaying outstanding performance activities they value. In the 1960s and 1970s, scholars in the field were emphasizing these valued strengths (e.g., Riessman, 1962; Mercer & Lewis, 1978; Torrance, 1973). Later, Baldwin (1985) promoted curriculum developed around areas of strength and interest for the purpose of developing thinking skills in another. Others have also advocated that diversity be valued and actively employed in programs (e.g., Florey & Tafoya, 1988; Maker, Nielson, & Rogers, 1994; Passow, 1986) and that strengths be emphasized during evaluations that are based on both test and nontest information (Frasier, 1991).

Given reported findings concerning the importance of teachers' beliefs in classroom practices (Hook & Rosenshine, 1979; Richardson, Anders, Tidwell, & Lloyd, 1991; Wilcox, 1982), cultural values may have an impact on what is recognized as worthy, who is seen as *talented,* and who is missed. Spindler and Spindler (1990) studied an idealistic, well-intentioned teacher as he interacted with students. His bias was "consistently in the direction of positive appraisals for upper status and mainstream children" (p. 65). The authors summarized as follows: "One's cultural background significantly influences what one will value, disvalue, and ignore.... (The teacher) simply did not interact in the same way or with the same intensity with children who did not match his own cultural experience and background" (p. 68). Cicourel and Kitsuse (1963) found, in regard to tracking for college, that school personnel did not apply objective criteria consistently in making decisions about students (pp. 63-65). Implicit, as well as explicit, factors were involved, including social class.

When mainstream teachers, unwittingly reflecting dominant culture values in their criteria, make recommendations for programs for children with high ability, those who are missed may be highly capable students whose culture does not value and promote verbal assertiveness around authority, display of knowledge for teachers or strangers or on standardized tests, factual knowledge (Bands, 1989; Garrison, 1989), or conspicuous and competitive individual achievement (cf. Spindler & Spindler, 1990). Those whose style of interaction is not mainstream (Ford, 1996; Goodwin, 1990; Phelps, Meara, Davis, & Patton, 1991), who have limited English proficiency (Maker, 1989), or who do not show obvious interest in subject matter deemed important by teachers may also be missed. Nonmainstream students may be deemed *unsuccessful* and *not a good fit* in programs designed for high achievers in the traditional curriculum (Barkan & Bernal, 1991; Maker, 1996) or may be "counseled out" (Maker, 1989, p. 294). Giftedness may, in fact, be recognized in terms of assimilation into the dominant culture, with "disadvantaged" children needing to distance themselves from their culture of origin in order to be *gifted* (Margolin, 1994).

Capable students with a low level of acculturation into the mainstream may find it difficult to participate assertively in groups or in other classroom interaction (Lake, 1990). In addition, talented students who are tired, poorly nourished, or distracted by family concerns might not participate enough in classroom activities to be recognized and recommended (Peterson, 1997). Richert (1987) noted that a large part of the inequity regarding identification is a function of socioeconomic level. According to *National Excellence* (1993), only 9 percent of students participating in programs for the talented and gifted were in the bottom quartile of family income as compared with 47 percent in the top quartile.

Much discussion has been focused on identification procedures, but not on the sociology of classroom interaction during those processes, particularly the role of teachers and cultural values. Attention to values across several groups, simultaneously assessing value orientation for each, is also lacking in the literature. Tannenbaum (1990) maintained that more than the "tried and false solution" of improved assessment instruments needs examination: "After so many years of nibbling futilely at the edges of these issues, it is time to bite boldly into their bitter core" (p. 84). Addressing one core issue, Borland (1993) argued for a postpositivist perspective on matters related to gifted education, emphasizing context and social construction in regard to conceptualizations. The studies of interest here, employing postpositivistic methods, addressed the issue of underrepresentation of nonmainstream children in programs, focusing on values across several cultural groups and their potential impact on selection processes.

The Original Study

In the first study (Peterson & Margolin, 1997), ethnographic analysis of the language of mainstream (see "Author Note" for a definition of mainstream), Midwestern middle school teachers in Diluvian revealed that assertiveness and classroom contribution were significant informal criteria when they made referrals for gifted programs. Most of the teachers' definitions of giftedness required verbal or other assertiveness and *showing* something — eagerness, being interested, creativity, being willing to share, a strong knowledge base, a sense of humor, and so forth.

An anthropological framework was used to view the 42 different conceptualizations of giftedness that emerged during analysis of the dominant-culture teachers' language. The teachers' definitions appeared to reflect the mainstream, dominant-culture value of individual, conspicuous, competitive achievement (Spindler & Spindler, 1990), a value with potential impact on the nomination of nonmainstream students. Students were nominated, for example, for a strong work ethic, motivation toward individual achievement, goal orientation, "standing out," and a competitive display of talents. A comfortable and positive relationship with the teacher was also often cited. In general, behavior was mentioned most often during nominations. Verbal ability, verbal assertiveness, and family status were noted almost as often, followed by a strong work ethic and social skills. Oddly, intelligence was mentioned only rarely. Almost all of these most-mentioned criteria are conspicuously absent in Marland's (1972) federal definition.

That the number of nominations for Latino children, the most significant nonmainstream group in the schools, was considerably lower than their actual demographic distribution was not surprising: Underrepresentation of nonmainstream children is well documented *(National Excellence,* 1993, p. 16*)*. The teachers' ad hoc definitions, however, offer a partial explanation of why such underrepresentation persists, in spite of concerted efforts to correct it, through more culture-fair tests, inclusive philosophies, and the use of multiple criteria for selection, most of which were being employed in that school district. Teachers' attitudes and opinions about giftedness and gifted programs and their personal feelings about individual students played significant roles in determining eligibility. Nominees were students who generally helped teachers to feel good about their work. Most important, giftedness was discussed with an apparent assumption that it was conceptualized similarly across all contexts. Of the 55 who participated, only one teacher asked, "By whose definition?"

The Extension of the Study

That study was extended to the Latino community in Diluvian and eventually involved four other nonmainstream communities elsewhere, including a low-income, mostly Anglo community. The purpose was to explore the possibility that the various nonmainstream populations would conceptualize giftedness differently from each other and differently from mainstream teachers. Any illumination of those differences might offer support for the finding in the original study that ad hoc selection criteria applied by teachers during the referral process reflected mainstream, dominant-culture values, and that application of these criteria might preclude nomination of nonmainstream children for gifted programs. Differences emerging between and among groups might provide additional evidence that mainstream teachers might not recognize talents and strengths that are highly valued in nonmainstream cultures and that they might insist on gifted behaviors that actually counter significant nonmainstream cultural values. Values of minority cultures might also constrain classroom behaviors valued by mainstream teachers. In addition, although exceptional abilities might be present, including those listed in program criteria, they may be demonstrated only within the nonmainstream language and culture or demonstrated differently from what is typical in the mainstream. The emphasis on behavior in the language of the nominating teachers in the original study suggests that perceptions about differing behaviors are critical in the nomination process.

Method

A Postpositivistic Mode of Inquiry

The postpositivist paradigm challenges the general emphases of positivism: objective reality, separation between researcher and subject, generalization, linear causality and prediction, and value-free inquiry. These are "based upon invalid notions about the nature of reality and inquiry" (Borland, 1990, p. 162). Postpositivists assume that "everything influences everything else" (Lincoln &

Guba, 1985, p. 151), including researcher and subject: "Both researcher and researched must have input into the inquiry process; and both must be open to being 'informed and transformed' as a result of the inquiry" (Guba, 1990, p. 88). Postpositivists are concerned with understanding rather than prediction; they are speculative, not deterministic. Given the importance of context in lives and events, generalization "is both unattainable and undesirable" (Borland, p. 163).

The postpositivist mode of inquiry has roots in social science research traditions in which educational research has linked schooling and social stratification (e.g., Bennett & LeCompte, 1990; Bourdieu & Passeron, 1977; Lareau, 1989), including how teachers' attitudes influence outcomes. One question asked by educational sociologists is how class and ethnicity interact with schooling to create stratification in society (Lancy, 1993). Coleman, Sanders, and Cross (1997) noted that "researchers [in the transformative mode] strive to make apparent those relationships so that people can understand their place in the world and be able to transform it" (p. 107).

In a postpositivistic mode, researchers can examine methods people use to create rules, to *account,* to convince, and to come to decisions to accomplish a reality (Mehan & Wood, 1975). The researcher can ask people to "think out loud" (Lave, 1988, p. 160): "By asking for accounts of the taken-for-granted from informants, (the researcher) insures that what is produced are rule-like guides for the uninitiated" (p. 185). With interest in the production of social structures, this methodology can be applied to the rules-creating processes of classroom teachers or nonmainstream community members as they identify individuals as gifted, labeling them as deviant, if giftedness is indeed conceptualized as an extreme in a normal distribution. The focus, then, shifts from examining how giftedness is manifested within an individual to the processes by which community members apply the label in *accomplishing giftedness.*

A postpositivist view of giftedness includes that it is "different things to different people... a social construction" (Borland, 1993, p. 12) — a different reality in each beholder's eye. In this phenomenological view, there is no *true* definition that can be *found* and no *accurate findings* of *gifted children.* Rather, giftedness is constructed by each person, depending on understandings brought to bear on that person's conceptualization, including cultural factors. According to the review of literature, there has long been discomfort with prevailing definitions and procedures for *locating gifted children.* The studies that are the focus here may help to explain that discomfort in displaying how differently a number of representatives from both mainstream and nonmainstream groups conceptualized giftedness.

Application to the Studies of Interest

Assuming, then, that giftedness is not a single objective reality (cf. Lincoln & Guba, 1985), a postpositivistic mode of inquiry was used here to explore ethnographically the *construction* of giftedness. A participant-observer researcher was an interactive instrument, with strategies meant to foster "dialogue and democratic theory building" (Roman & Apple, 1990, p. 63). The sampling was intentionally nonrandom and purposive (Lincoln & Guba,1985). Appropriate to ethnographic inquiry, sites and participants were small in number (Table 1). The data were words, and narrative responses were the focus of phenomenological analysis. The investigator was aware of her biases concerning classroom teachers' referrals and levels of multicultural awareness and the importance of nonmainstream voices as she attempted to capture the subjective reality of the participants. Hypotheses emerged from the data (cf. Lancy, 1993). The researcher subsequently conducted member checks by presenting findings either orally (e.g., inservice presentations to teachers) or in writing (e.g., narrative summaries) to either individuals or groups in each community in order to check out resonance of interpretive accounts with participants' actual experience. All of these aspects are related to "theoretical and political adequacy" (Roman & Apple, p. 63).

With this mode of inquiry, setting is important; and the researcher is immersed in it. The settings in these studies were in each group's natural environment: schools, homes, churches, a community center, and a tribal center. The process involved "thick description" (Geertz, 1973): an emphasis on particulars of language, activity, and context. Intuitive knowledge was important in understanding nuances of interaction and multiple realities of respondents (cf. Borland, 1990).

Basic to postpositivistic inquiry is the concept of grounded theory, "the discovery of theory from data systematically obtained from social research" (Glaser & Strauss, 1967, p. 3), as opposed to theory generated from a priori assumptions. The design also emerges from the interaction. During the first study, the emphasis on cultural values was precipitated by the themes that emerged in the teachers' language; and the theory regarding nomination of nonmainstream children arose from speculation about differences between teacher and student value orientations. During the second study, clear value differences that emerged in the language of the five nonmainstream groups — and between teacher values and nonmainstream values — led to theory concerning why nonmainstream children are underrepresented in gifted programs.

Participants

Primary participants in the extension of the study were individuals from various nonmainstream populations in several communities located within one midwestern state: Latinos in Diluvian, Immigrant Asians in Archway, African Americans in Central City, American Indians in an Indian settlement, and low-income Anglos in Armbauer. Dominant-culture teachers were also interviewed in Central City and at the Indian settlement. (See Table 1 for demographic information.)

Procedures

Most of the data for this study was gathered through interaction with individuals or small groups from the mainstream and nonmainstream populations. A small portion of the data was solicited with a writing prompt from students in family science classes in a low-socioeconomic community. (Table 1 includes a summary of interviewing arrangements.) The oral interviews were audiotaped and, in several cases, were recorded simultaneously on laptop computer. Field notes were recorded after each session. Each interview or writing prompt focused on one question, either "Who have you known personally who is or was 'gifted'?" or "Who would you nominate as 'gifted' from your class?" Regardless of circumstances and setting, the basic objective remained constant: to generate language from teachers and nonmainstream individuals concerning who they perceived to be gifted.

Table 1

Major Themes in the Language of Five Nonmainstream Groups in Response to "Who Have You Known Who is 'Gifted'?"

	Cultural Group	
	Latino	**African American**
Community	Diluvian; population: 23,000; 13% Latino	Central City; population: 300,000; schools 54% & 38% African Amerian
N =	12	22
Duration	3 weeks	6 weeks
Location	Community cultural center	2 schools: grades K-2, 3-5
Length of interviews	1 hour	30-120 minutes
Type of interview	Individual, Group	Individual, Group
Participants	Adults, youth	Administrators, aides, parents, grandparents
Major Themes	Arts (as expression, not achievement); humility; community service; helping extended family; loyalty to community	Contribution to neighborhood; handiwork; concern for family; wisdom (as distinguished from "book knowledge")

Participants' comments were subsequently analyzed for themes according to methods suggested by Glaser and Strauss (1967) and Altheide (1987) for ethnographic analysis. Data were not fitted into predefined categories. Instead, statements were coded according to emerging themes, and statements similarly coded were continually and reflexively compared with each other, with adjustments of the coding scheme often occurring. The data dictated interpretational routes in this largely inductive process. However, theoretical preconceptions also offered guidance, namely, that cultures differ in what is valued, that what is seen as gifted reflects what is valued in a particular context, that cultural differences mean that *gifts* are not recognized universally across cultures, and that mainstream teachers may not recognize that their conceptions of giftedness may preclude nomination of nonmainstream children.

American Indian	**Immigrant Asian**	**Low Socio-economic Anglo**
Settlement; population: 650; 100% American Indian	Archway; population: 2,900; 27% minority; 4% Asian	Armbauer; population: 6,000; 43% low income
11	12	39
2 occasions	1 meeting	2 meetings; 1 class
Tribal center, school	Church	Dept. of Human Services, high school
1-2 hours	3 hours	2-4 hours with adults; 1 school class period
Group	Group	Group
Tribal leaders, culture teachers	Adults, youth in U.S. 1-5 years	Adults, high school students
Ability in native language; ability to find satisfaction in both cultures; contribution to tribe; talent in music, art, dance, knowledge of culture	Education; ability to adapt; asceticism; hard work; focus on children; attention to tradition; respect for elderly; respect for the past	Helping others; child rearing; handiwork; manual dexterity; artistic talent; ability to overcome adversity; personality, social effectiveness; nonbook knowledge

Findings

Latinos

The Latino participants nominated mostly family members. The main themes that emerged were artistic talent, humility (as contrasted with competitive self-assertion), and community service (but not through organized activity, such as school activities or Scouts). Every Latino participant cited artistic talent in at least one nomination, and that area of talent prompted the largest quantity of language. Of particular interest to this study was the observation in the original study that mainstream art teachers invariably emphasized perfection, detail, and skills; the Latinos here focused exclusively on art as an avenue of expression. The following are examples of comments representing two of the main themes:

Artistic talent

She expresses herself in drawing. If she's sad, it gets her to feel better.

He lets the drawing express itself, and he expresses himself through that.

Humility

My parents taught us to work hard, but not be show-offs. You shouldn't put yourself above anyone else. If we have more than others, then we should help to make them be equal. My dad said we should try to get ahead. If we succeed, good, but don't get a big head.

African Americans

As with the Latinos, almost all individuals nominated by African Americans were close to home. The main themes that emerged in their language were as follows:

Selfless contribution to the neighborhood (e.g., being active in church; serving on boards; nurturing children; doing nice, thoughtful things)

She works tireless to help the community. (founder of a neighborhood daycare center)

He had a dream of making a home for older people in the inner city. (pastor)

He helps the school get their money for tutoring the children. (husband)

Out of the goodness of her heart she takes kids in the neighborhood, and she will do their hair and just make them feel good about themselves, a real role model. (neighbor)

Handiwork

He doesn't have to read books to do this. Can just build a deck and doesn't need instructions. His home is just beautiful. He's very good at fixing things, gifted with his hands. Built a hotrod when he was a kid and won trophies. (brother)

He can make something simple into a work of art. (husband)

Concern for family

In terms of his love for his family, the tutelage he gave his family. (father)

Stepped in for the mother, takes him shopping, reads books to him, has picked up frogs with him. That's an aunt he'll always remember, a very special person. (daughter)

Wisdom

The wisest man I've ever encountered in my life — his desire to educate himself, so knowledgeable about so many subjects, and without a great deal of formal education. (father)

Unique to this group was the emphasis on contribution to the neighborhood through, for example, involvement in church, serving on boards, and assisting others.

Immigrant Asians

Major themes in the interaction with the immigrant Asians were as follows:

Education

She goes to college, and she's going to go again for another degree.

Adaptation, change

When he came, he had just his hands. He had nothing. After two or three years, he saved some money and rented a room and opened with two to three tables for a restaurant. He cooking everything good. Four or five years so successful he bought it.

Caring for family

Anyone who can take care of the family.

Asceticism and hard work for the future

My uncle. He ate only one time a day. He cut trees for the wood, and he sells it and kills squirrels and snakes and sold it at the market. And he doesn't buy anything at the market and eats only what he gets and only buys rice at the market. Makes things out of wood. He just liked living alone. He went to the temple to become a monk. I admire it because he needs so little.

Noteworthy in this group was a minor theme that appeared several times in connection with individuals who had succeeded through effort in the United States and who had businesses as evidence of success. There was repeated emphasis that noteworthy individuals in the lands of their heritage might rather be non-materially-oriented minimalists, with gifts of loyalty to family and culture, a sense of tradition, and solid character.

American Indians

The tribal leaders at the Indian settlement were unwilling to name anyone as gifted. One explained, "You don't put yourself above anyone." Another said, "The idea of helping — we work together. The projects — when we get through, it's not 'my project.'" Someone else summarized: "We're taught not to put ourselves above others. Obviously there are people who have come to the forefront, doing good for people, providing leadership. They do it quietly." It was pointed out that the culture does not value bragging. Using a word like *gifted* suggests that someone is better, more valued than others, set above others. The idea of giftedness is also difficult because it is seen as assessment, something not promoted in the culture. The interview moved instead to a reluctant discussion of what qualities were respected, and the following statements are representative: "Educated, knowledgeable, traditional, practices the culture, participates in ceremonies, can blend the cultures, can find satisfaction in both, without becoming assimilated." "Taking an active role in monitoring the changes in the tribe so the culture is still intact for the next generation-to do as much as an individual can do to contribute." The local Indian culture teachers responded similarly, naming no one, but offering comments about what is respected:

Can absorb, can catch the language real fast
Are talented in music and dance, singing
Are knowledgeable about the culture
Are able to perform and reflect through both cultures

References to motivation for individual achievement, to the work ethic, to factual knowledge, or to advanced-level work were noticeably absent. Creativity was a basic theme for all who were interviewed on the settlement, whether member of the tribe or dominant-culture teacher. It should be noted that the settlement school had no gifted program.

Low-Income Anglo Americans

When the low-income Anglo American adults responded to the question about whom they had known personally who was gifted, there was no reference to success in terms of status or material possessions. Rather, gifted persons were nurturers, listeners, helpers, and advisors — "being there" for others. Giftedness was also broadly conceptualized as "being able to do what you enjoy doing" or "doing with what you were given." Nominees were versatile, had practical skills, were wise, and were storytellers. Most had overcome adversity.

More than in any of the other communities, here the participants asked for clarification of the word *gifted* and emphasized that "it could be anything." One said that it could be to "do bicycles, cars, bricks." They articulated that there might be differences between their views and those of someone from a different geographical area. They nominated a total of 34 individuals, most being members of extended family, but also doctors, teachers, and a pastor. Their major themes follow, in rank order, from most to least:

Helping others, listening, advising

He was the one people turned to.

Child-rearing, teaching

Tells them the right thing.

Manual dexterity, creativity, success in outdoor enterprises

Nothing he couldn't do on a house.

Calm hands, like a surgeon.

Makes beautiful things

Academic ability, with practical application

Gifted at math — did all the accounting for my lawn business, saved me some money.

Overcoming adversity

My dad was killed, so my mom had to raise both of us on her own. Just finished a secretarial degree last year. She never let that feeling of being alone interfere with us.

He has dyslexia, but he's gotten around that.

Using the writing prompt, the family science teacher assigned an essay to entire classes; then, according to approved project format, selected out the essays of those on reduced lunches. The following themes emerged, most of them similar to the themes found in the adult groups at Armbauer:

- Ability to listen, understand
- Ability to overcome adversity
- Ability to nurture children
- Artistic ability, especially drawing
- Athletic ability
- Academic ability, school skills, writing ability
- Being virtuous — honest, trustworthy, patient, caring

Only writing ability and athletic ability from the above were not mentioned by the adults in their community, those two themes perhaps reflecting the school context of the adolescents.

Again, resistance to using the word *gifted* was strong in these essays, although, unlike at the Indian settlement, here the students also challenged how it was conceptualized, not just that it was being discussed. The student essays included statements like these: "I don't feel like anybody is more gifted than anyone else. Some people might show it more than others, and some might not know they're gifted or just don't want to show it." "Gifted is a general word. Being more specific would help. Like being gifted in a certain area such as sports or school." In this group of adolescents, all but five of the nominees were in the students' extended families.

Common Themes Among Nonmainstream Groups

A comparison of the themes that emerged in the language of all nonmainstream groups found that the following themes appeared in three or four of the groups:

- Helping others, informal networking, contribution to community, listening (4)
- Manual dexterity, artistic ability (4)
- Concern for family, children (4)
- Social effectiveness, leadership *(3)*
- Nonbook knowledge *(3)*

Two of the groups valued *not* displaying what one knows.

An additional theme in the language of all nonmainstream groups was appreciation for being asked for opinions and for having an attentive listener from the dominant culture. During one interview with two African American women, one said, "I think we think our opinion doesn't count as much. That there won't be many of us there anyway, like at the [parent] meeting, so it doesn't matter." The other said this:

> At the meeting, most of those people are much more educated than me, but I listened to what they were saying. I was noticing the words they used to describe their children. I noticed that they talked so distinguished. They have the words, the education. They use language that way. If you're not into that, if you don't have the language, you can't do that. The Caucasian people who spoke — I wouldn't even think of those words to use.

Feedback From the Communities

During subsequent presentations to cooperating school districts, including to school administrators and gifted-education personnel, educators indicated receptivity to the idea of differing cultural values and to implications for selection processes. Several mainstream teachers commented privately that they "hadn't thought about things like that before," their comments echoing those of several nonmainstream individuals as well. The following comment from an American Indian was pessimistic, however:

> They know only the white view. They have never had to think of another view. The white culture doesn't really care to sort that out. They don't take the time. They make generalizations. They'll start to get conflicting messages. Whites don't like that.

Another settlement resident commented, "The dominant culture isn't interested in learning about us." A tribal leader observed, "We understand and appreciate the white culture more than they understand us." A Latino mother emphasized the need to adjust to the majority culture: "We need to educate the parents to know that it is important for their children to be involved at school and how children should behave so that they will be recognized."

Dominant-Culture Teachers at the Indian Settlement and in Central City

When the all-Anglo teachers at the settlement school were interviewed, the new and veteran (on staff more than three years) teachers were interviewed separately, after the Anglo principal indicated a desire to see if there was a difference between the groups. The newer teachers listed children by name and mentioned exceptionality in such areas as creative thinking, math, art, confidence, and curiosity. However, their language often had a *yes, but* quality, as if they could not affirm a strength of an Indian child without indicating something that challenged or qualified it (e.g., "creative thinking, but not very good reading or writing" and "artistic, but trouble with reasoning, math"). Curiously, all work-ethic references and most of the references to verbal expression, both reflecting mainstream cultural values, were made with no *but* qualifier.

The veteran teachers' language included no *but* phrases, and they mentioned few children by name. Although a few of their themes were similar to those of teachers in other settings in the study (e.g., creativity, eagerness to learn, advanced-level work), most were not (e.g., storytelling, consciousness of family, formulating answers before speaking), reflecting an apparent sensitivity to the culture. When one veteran teacher was asked for her views on the differences between the two teacher groups, she avoided simple notions related to tenure in the school and hypothesized that the veterans were fairly homogeneous regarding rural upbringing, large families of origin, and low socioeconomic status. Their backgrounds contrasted with those of the newer teachers, most of whom came from middle-class, urban backgrounds and smaller families. The teacher speculated that teachers with backgrounds more like families in the Indian settlement tended to remain and become more and more acquainted with the local culture.

Teachers in the two Central City school were interviewed in small groups according to grade level. They had had in-service training regarding Gardner's (1983) theory of multiple intelligences and regularly nominated children in the four areas promoted in the district (language, math, leadership, and art). At the end of each interview, the teachers were asked for a "bottom-line definition" of giftedness. Here their language abruptly reverted to dominant-culture values. Table 2 provides a comparison of mainstream and non-mainstream themes.

Discussion

The teachers in the original study conceptualized giftedness as being what is *good* in students, that is, what is apparently valued within the school culture. Similarly, individuals representing nonmainstream cultural groups appeared to

Table 2

Comparison of Common Themes in the Language of Individuals From Five Nonmainstream Communities and Themes in the Language of Dominant-Culture Teachers in Two Nonmainstream Communities

Major themes of Diluvian teachers in original study	Themes of dominant-culture teachers in Central City when nominating children as "gifted" with multiple intelligences as a guide	Themes of dominant-culture teachers in Central City when giving "bottom line" regarding "giftedness"	Common themes among individuals from five nonmainstream communities nominating "gifted persons"
Behavior	Verbal skills	Individualism and individual achievement	Listening, helping others, contributing to community
Verbal ability, verbal assertiveness	Creative, artistic talent	Conspicuous achievement	Manual dexterity, artistry
Family status	Leadership	Work ethic, motivation	Concern for family, children
Work ethic, motivation	Work ethic, motivation, task commitment	Assertiveness	Social effectiveness, leadership
Social skills	Behavior		Nonbookish learning

communicate what they valued when nominating gifted persons. Given the remarkable differences in what emerged as valued attributes and behaviors in the language of mainstream classroom teachers and nonmainstream groups, it is possible that members of each group might be missed by the other for a gifted program if nomination were based on what each considered valuable. Teachers might not refer children who adhere to traditional minority-culture values since, according to this study, the values of nonmainstream groups apparently often do not encourage individual, competitive, conspicuous achievement and assertive, self-promoting display of talents. However, nonmainstream individuals might also not nominate mainstream individuals because they do not embody, for example, humility, collaboration, selfless concern for family, altruistic service to the neighborhood, manual dexterity, ability to listen and guide, and wisdom. Gifted programs in each of the nonmainstream groups might look quite different from those based on valued mainstream behaviors and qualities: The Latinos might promote expressive arts and collaborative activity; the African Americans might select individuals based on selfless contribution; the immigrant Asians might use respect for the family and ability to adapt as criteria; low-income Anglo Americans might focus on nonbookish learning and the ability to listen to others. The American Indians would not be likely to create a selective program at all, certainly not focusing on individual accomplishments.

The concept of giftedness appears to be bound to context. Each culture sees *goodness* through its own cultural lens, including the dominant culture, which has its own particular value orientation. The *educator culture* of teachers in this study — of whom 98 percent were from the dominant culture, and who, by virtue of income level, were from the middle class — undoubtedly added its own idiosyncratic valuing system to their already dominant-culture value orientation. Yet educators, children and adolescents, and parents are probably not cognizant of the impact of their values on whether doors are opened or closed regarding special educational opportunities, those opportunities increasing in number with each advancement in grade level. Students who do not demonstrate individualistic, assertive, and competitive behaviors may, in fact, be seen as *less,* in an assumed hierarchy of values. They may not be seen as the best, when the best are nominated for scholarships, leadership experiences, special positions on teams or committees, or gifted programs. Nonmainstrearn parents might not know how educators measure *worth*, and educators might not understand that themselves. To understand would require self-examination, and the dominant culture is usually not required to examine itself as a culture, since its values are often the standard by which others are measured. Educators may not recognize that considerable comfort in the classroom is likely a prerequisite for many or most of the gifted behaviors they are looking for and that cultural differences may inhibit or even preclude those behaviors.

In the complex, mobile mainstream society of today, many families are distant from extended family and feel isolated within the individualistic, competitive, conspicuous-consumption culture that surrounds them. It may be increasingly important, therefore, to value networking, collaboration, nonjudgmental listening, humility, family loyalty and support, the wisdom of age, expressive arts, respect for authority, automatic deference to teachers, and a deemphasis on the trappings of social status, which are several of the many themes that emerged in the language of the nonmainstream groups in this study. These themes contrast with some of the primary values of the dominant culture, as reflected in the language of the teachers in the study. There is much to learn from, not merely about, the nonmainstream cultures, based on the themes that appeared. Probably through any cultural lens, the values that emerged in their language are "good."

The findings here argue for vigorous staff development for educators regarding the impact of value orientations in the identification and selection for gifted programs and other selective processes; for creative approaches to affirming culturally valued gifts and talents in the classroom and in special programs; for employing teaching strategies that accommodate the cultural values of nonmainstream students; for involving community in the identification and selection process; and for educating parents about special programs, opportunities, and purposes and criteria for programs. The findings also suggest that gifted education, long an innovator regarding educational practices, might take the lead in developing a *cultural-lenses curriculum* that presents values nonhierarchically — respecting and valuing the values of various cultures — instead of merely tolerating them, as the tolerance language of multicultural awareness raising seems to encourage. Students might not only reflect on their own values but also analyze the values of their peers, encouraged to see them only as different from each other, not better or worse (cf. Montgomery, 1989). In short, rather than focusing mostly on refining methods for identifying students for existing programs, in concert with shifting paradigms, gifted education might also reexamine the existing programs.

As one Diluvian teacher said informally after a group interview, "We don't really understand about cultural differences." If she and her colleagues did understand them, they might appreciate and celebrate cultural uniquenesses, including the potential for a broad range of differences within each particular cultural group. They might be aware of the special strengths of minority cultures — and the gifts that might not be demonstrated in ways with which mainstream teachers are familiar. If gifted education personnel did understand, they might view fitting into the program differently, they might reconsider what are the good behaviors that constitute giftedness, they might reconsider individual needs in regard to program philosophy, they might develop more flexible curricula, they might consider enrichment in a different way, and they might be inspired to learn through listening and observing more about the nonmainstream cultures represented in their schools.

The majority culture is often quick to criticize minority groups for lacking dominant-culture-style high aspirations and motivation, two behaviors valued in the teacher culture, according to the original study. However, in addition to the fact that high aspirations is also a context-bound construct, it is important to recognize that aspirations might mean moving away from the culture of origin, from a stable support system, and into an individualistic way of life that may lack readily available support. Education might, therefore, be dangerous. What should minority individuals be willing to give up for the sake of high aspirations? The Anglo teachers at the Indian settlement spoke of the settlement "pulling kids back," out of the nonsettlement high school. The tribal leaders felt that if they had "to make a choice between cultures," the choice would be the home culture. Rather than indicting minority cultures, educators might ask these questions: Has the majority culture graduated from values, such as mutual support and interdependence, support for family, respect for tradition, emphasis on nonbookish wisdom and skills, and listening? Are these behaviors and values what the majority could learn from minority groups, with additional and mutual benefit through that process of learning? Could gifted education collectively recognize these gifts, according to nonmainstream value orientations, and incorporate them into their programs as well?

A postpositivistic mode of inquiry is appropriate for exploring these and other questions that are surfacing as paradigms shift regarding educating the gifted. In moving away from long-dominant modes of studying talent development (cf. Coleman et al., *1997),* researchers can explore complex and highly nuanced phenomena in school and community contexts that have an impact on basic program processes — and on students. Scholars can similarly study the students themselves, with altered conceptualizations a possible result. As a counselor educator, the researcher is familiar with basic counseling tenets, such as focusing on strengths, building a collaborative relationship, respectfully entering the world of clients, and encouraging them to teach the counselor about that world. These tenets have potential for useful application at many levels of education in a pluralistic society. Much of that phenomenological posture fits well with postpositivism and offers yet another cultural lens that can enhance understanding and contribute to theory building, not unlike the lenses of the nonmainstream groups in the study just described.

Author Note

Mainstream has no precise definition, but traditionally it has been considered to be Anglo-Saxon and Northern European in heritage and Protestant. If those from European Catholic stock are included, they constitute three-fourths of the population of the United States (Spindler & Spindler, *1990).*

References

Adderholdt-Elliot, M., Algozzine, K., Algozzine, B., & Haney, K. *(1991).* Current state practices in educating students who are gifted and talented. *Roeper Review, 14,* 20-23.

Altheide, D. L. *(*1987*).* Intelligence testing and Chicanos: A quality of life issue. *Social Problems, 27,* 186-195.

Baldwin, A. Y. (1978). Introduction. In A. Y. Baldwin, G. H. Gear, & L. J. Lucito (Eds.), *Educational planning for the gifted: Overcoming cultural, geographic and socioeconomic barriers* (pp. 1-5). Reston, VA: Council for Exceptional Children. (ERIC Document Reproduction No. ED 161 173)

Baldwin, A. Y. (1985, April). *Identification and programming for Black gifted children.* Paper presented at the Annual Convention of the Council for Exceptional Children, Anaheim, CA. (ERIC Document Reproduction No. ED 261 495)

Banda, C. (1989). Promoting pluralism and power. In C. J. Maker & S. W. Schiever (Eds.), *Critical issues in gifted education: Defensible programs for cultural and ethnic minorities* (pp. 27-33). Austin, TX: PRO-ED.

Barkan, J. H., & Bernal, E. M. (1991). Gifted education for bilingual and limited English proficient students.

Gifted Child Quarterly, 35, 144-147.

Bennett, K. P., & Lecompte, M. D. (1990). How schools work: Sociological analysis of education. White Plains, NY: Longman.

Bernal, E. M., Jr. (1974). *Gifted Mexican-American children: An ethnico-scientific perspective.* (ERIC Document Reproduction No. 091 411)

Bernal, E. M., Jr. (1977). Assessment procedures for Chicano children: The sad state of the art. *Aztlan-International Journal of Chicano Studies Research,* 8, 69-81.

Bernal, E. M., Jr. (1981). *Special problems and procedures for identifying minority gifted students.* (ERIC Document Reproduction No. ED 203 652)

Borland, J. H. (1990). Postpositivist inquiry: Implications of the "new philosophy of science" for the field of the education of the gifted. *Gifted Child Quarterly,* 34, 161-167.

Borland, J. H. (1993). Giftedness and "The new philosophy of science." *Understanding Our Gifted,* 5(6), 1, 11-14.

Bourdieu, P., & Passeron, J. (1977). *Reproduction in education, society, and culture.* London: Sage.

Bruch, C. B. & Curry, J. A. (1978). Personal Learnings: A current synthesis on the culturally different gifted. *Gifted Child Quarterly,* 22, 313-321.

Buchanan, N. K., Javier, L., Sing, D., & Plunkett, T. (1993). *Performance-based identification of culturally diverse gifted students: A Pilot Study.* Preliminary Report Prepared for The Pacific Rim Symposium on Higher Education Evaluation. (ERIC Document Reproduction Service No. ED 369 249)

Buckley, K. C. P. (1992). A survey of educational values and conceptions of gifted intelligence held by parents who have enrolled their children in programs for the gifted. *Dissertation Abstracts International,* 54(1), 127-A.

Callahan, C. M., Tomlinson, C. A., & Pizzat, P. (April, 1993). *Contexts for promise: Promising practices in the identification of gifted and talented students.* Charlottesville, VA: The National Research Center on the Gifted and Talented, University of Virginia. (ERIC Document Reproduction Service No. ED 372 592

Cicourel, A. V., & Kitsuse, J. I. (1963.) *The educational decision-makers.* Indianapolis, IN: Bobbs-Marrill.

Colangelo, N., & Parker, M. (1981). Values of gifted students. *Counseling & Values,* 26(1), 35-41.

Coleman, L. J., Sanders, M. D., & Cross, T. L. (1997). Perennial debates and tacit assumptions in the education of gifted children. *Gifted Child Quarterly,* 41(3), 105-111.

DeAvila, E. A., & Havassy, B. (1974). *IQ tests and minority children.* Austin, TX: Dissemination Center for Bilingual Bicultural Education. (ERIC Reproduction Document No. ED 109 261

Duran, B. J., & Weffer, R. E. (1992). Immigrants' aspirations, high school process, and academic outcomes. *American Educational Research journal,* 29, 163-181.

Feldhusen, J. F. (1998). A conception of talent and talent development. In R. C. Friedman & K. B. Rogers (Eds.), *Talent in context: Historical and social perspectives on giftedness* (pp. 193-209). Washington, DC: American Psychological Association.

Feldman, D. H. (1992). Has there been a paradigm shift in gifted education? In N. Colangelo, S. G. Assouline, & D. L. Ambroson (Eds.), *Talent development: Proceedings from the 1991 Henry B. and Jocelyn Wallace National Research Symposium on Talent Development* (pp. 89 94). Boston: Trillium Press.

Florey, J., & Tafoya, N. (1988). *Identifying gifted and talented American Indian Students: An overview.* Las Cruces, NM: ERIC Clearinghouse on Rural Education and Small Schools. (ERIC Document Reproduction No. ED 296 810)

Ford, D. Y. (1996). *Reversing underachievement among gifted Black students: Promising practices and programs.* New York: Teachers College Press.

Frasier, M. M. (1991). Disadvantaged and culturally diverse gifted students. *Journal for the Education of the Gifted,* 14, 234-245.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* New York: Basic Books.

Garrison, L. (1989). Programming for the gifted American Indian student. In C. J. Maker & S. W. Schiever (Eds.), *Critical issues in gifted education: Defensible programs for cultural and ethnic minorities* (pp. 116-127). Austin, TX: PRO-ED.

Geertz, C. (1973). Thick description: Toward an interpretative theory of culture. In C. Geertz (Ed.), *The interpretation of cultures: Selected essays by Clifford Geertz* (pp. 3-32). New York: Basic Books.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago: Aldine.

Goodwin, M. H. (1990). *He-said-she-said: Talk as social organization among Black children.* Bloomington: Indiana University Press.

Grant, B. (1995). The place of achievement in the life of the spirit and the education of gifted students. *Roeper Review, 18*, 132-134.

Guba, E. G. (1990). Subjectivity and objectivity. In E. W. Eisner & A. Peshkin (Eds.), *Qualitative inquiry in education: The continuing debate* (pp. 74-91). New York: Teachers College Press.

Hoge, R. *(1988).* Issues in the definition and measurement of the giftedness construct. *Educational Researcher, 43,* 12-16.

Hook, C. M., & Rosenshine, B. V. (1979). Accuracy of teacher reports of their classroom behavior. *Review of Education,* 49(2), 1-12.

Howard-Hamilton, M., & Franks, B. A. (1995). Gifted adolescents: Psychology, behaviors, values, and developmental implications. *Roeper Review, 17,* 186-191.

Kingore, B. W. (1990). *Kingore observation inventory (KOI).* Des Moines, IA: Leadership Publications.

Kleinsasser, A. (1986). *Exploration of an ambitious culture: Conflicts facing gifted females in rural communities.* Paper presented at the Annual Conference of the National Rural and Small Schools Consortium, Bellingham, WA. (ERIC Document Reproduction No. ED 278 522*)*

Lake, R. *(*1990, September). An Indian father's plea. *Teacher Magazine,* 51-53.

Lancy, D. F. (1993). *Qualitative research in education: An introduction to the major traditions.* New York: Longman.

Lareau, A. (1989). *Home advantage: Social class and parental intervention in elementary education.* New York: Falmer.

Lave, J. (1988). *Cognition in practice.* New York: Cambridge University Press.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry.* Beverly Hills, CA: Sage.

Lubinski, D., Schmidt, D. B., & Benbow, C. P. (1996). A 20-year stability analysis of the study of values for intellectually gifted individuals from adolescence to adulthood. *Journal of Applied Psychology, 81,* 443-451.

Maker, C. J. (1989). Programs for gifted minority students: A synthesis of perspectives. In C. J. Maker & S. W. Schiever (Eds.),. *Critical issues in gifted education: Defensible programs for cultural and ethnic minorities (pp. 293-309).* Austin, TX: PRO-ED.

Maker, C. J. *(1996).* Identification of gifted minority students: A national problem, needed changes and a promising solution. *Gifted Child Quarterly, 40,* 41-50.

Maker, C. J., Nielson, A. B., & Rogers, J. A. *(1994).* Giftedness, diversity, and problem-solving. *Teaching Exceptional Children. 27(1),* 4-19.

Margolin, L. (1994). *Goodness personified: The emergence of gifted children.* Hawthorne, NY: Aldine DeGruyter.

Marland, S. J. (1972). *Education of the gifted and talented.* (Report to the Congress of the United States by the U.S. Commissioner of Education). Washington, DC: U.S. Government Printing Office. (ERIC Document Reproduction Service No. ED 056 243)

McKenney, N. R., & Bennett, C. E. (1994). Issues regarding data on race and ethnicity: The census bureau experience. *Public Health Reports,* 109(1), 16-25.

Mehan, H., & Wood, H. (1975). *The reality of ethnomethodology.* New York: John Wiley and Sons.

Mercer, J. R., & Lewis, J. F. *(1978).* Using the system of Multicultural Pluralistic Assessment (SOMPA) to identify the gifted minority child. In A. Y. Baldwin, G. H. Gear, & L. J. Lucito (Eds.), *Educational planning for the gifted: Overcoming cultural, geographic and socioeconomic barriers* (pp. 7-13). Reston, VA: The Council for Exceptional Children.

Mistry, J., & Rogoff, B. (1985). A cultural perspective on the development of talent. In F. D. Horowitz & M. O'Brien (Eds.), *The gifted and talented: Developmental perspectives* (pp. 125-144). Washington, DC: American Psychological Association.

Montgomery, D. (1989). Identification of giftedness among American Indian people. In C. J. Maker & S. W. Schiever (Eds.), *Critical issues in gifted education: Defensible programs for cultural and ethnic minorities* (pp. 79-90). Austin, TX: PRO-ED.

National excellence: A case for developing America's talent. (1993). Washington, DC: U.S. Government Printing Office.

Passow, A. H. (1972). The gifted and the disadvantaged. *National Elementary Principal, 51(3),* 24-31.

Passow, A. H. (1986). Educational programs for minority/disadvantaged gifted students. In L. Kanevsky (Ed.), *Issues in gifted education (pp.* 147-172). San Diego, CA: San Diego City Schools GATE Program.

Passow, A. H., & Frasier, M. M. (1996). Toward improving identification of talent potential among minority and disadvantaged students. *Roeper Review, 19,* 198-202.

Peterson, J. S. (1997). Bright, tough, and resilient, and not in a gifted program. *Journal of Secondary Gifted Education, 8,* 121-136.

Peterson, J. S., & Margolin, L. (1997). Naming gifted children: An example of unintended "reproduction." *Journal for the Education of the Gifted, 21,* 82-100.

Phelps, R. E., Meara, N. M., Davis, K. L., & Patton, M. J. (1991). Blacks' and Whites' perceptions of verbal aggression. *Journal of Counseling and Development, 69,* 345-350.

Reid, N. (1992). *Correcting cultural myopia: The discovery and nurturance of the culturally-different gifted and talented in New Zealand.* Wellington, NZ: New Zealand Council for Educational Research. (ERIC Document Reproduction Service No. ED 357 532)

Renzulli, J. S., & Delcourt, M. A. (1986). The legacy and logic of research on the identification of gifted persons. *Gifted Child Quarterly* 30, 20-23.

Renzulli, J. S., & Hartman, R. K. (1971). Scale for rating behavioral characteristics of superior students. *Exceptional Children,* 38, 243-248.

Richardson, V., Anders, P., Tidwell, D., & Lloyd, C. (1991). The relationship between teachers' beliefs and practices in reading comprehension instruction. *American Educational Research Journal,* 23, 559-586.

Richert, E. S. (1985). Identification of gifted children in the United States: The need for pluralistic assessment. *Roeper Review,* 8, 68-72.

Richert, E. S. (1987). Rampant problems and promising practices in the identification of disadvantaged gifted students. *Gifted Child Quarterly,* 31, 149-154.

Richert, E. S. (1992). *Equitable identification of students with gifted potential.* Washington, DC: Department of Education. (ERIC Document Reproduction Service No. ED 366 159)

Richert, E. S. (1997). Excellence with equity in identification and programming. In N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (2nd ed., pp. 75-88). Boston: Allyn and Bacon.

Riessman, F. (1962). *The culturally disadvantaged child.* New York: Harper & Row.

Roman, L. G., & Apple, M. W. (1990). Is naturalism a move away from positivism? Materialist and feminist approaches to subjectivity in ethnographic research. In E. W. Eisner & A. Peshkin (Eds.), *Qualitative inquiry in education: The continuing debate* (pp. 38-95). New York: Teachers College Press.

Scott, M. S., Perou, R., Urbano, R., Hogan, A., & Gold, S. (1992). The identification of giftedness: A comparison of White, Hispanic, and Black families. *Gifted Child Quarterly,* 36, 131-139.

Shen, W., & Mo, W. (1990). *Reaching out to their culture — Building community with Asian-American families.* (ERIC Reproduction Document No. ED 351435)

Spindler, G., & Spindler, L. (1990). *The American cultural dialogue and its transmission.* London: Falmer Press.

Tannenbaum, A. J. (1990). Defensible? Venerable? Vulnerable? An essay review of Maker and Schiever, *Defensible Programs for Cultural and Ethnic Minorities. Gifted Child Quarterly, 34,* 84-86.

Tonemah, S. A. (1991). Philosophical perspectives of gifted and talented American Indian education. *Journal of American Indian Education,* 31(1), 3-9.

Torrance, E. P. (1973). What gifted disadvantaged children can teach their teachers. *Gifted Child Quarterly,* 17, 243-249.

Torrance, E. P. (1977). *Discovery and nurturance of giftedness in the culturally different.* Reston, VA: Council for Exceptional Children. (ERIC Document Reproduction No. ED 145 621)

Torrance, E. P. (1978). Ways of discovering gifted Black Children. In A. Y. Baldwin, G. H. Gear, & L. J. Lucito (Eds.), *Educational planning for the gifted: Overcoming cultural, geographic and socioeconomic barriers* (pp. 29-33). Reston, VA: The Council for Exceptional Children.

Treffinger, D. J., & Feldhusen, J. F. (1996). Talent recognition and development: Successor to gifted education. *Journal for the Education of the Gifted,* 19, 181-193.

VanTassel-Baska, J. (1991). Cultural diversity in gifted education. *Understanding Our Gifted,* 3(6), 1, 10-11.

Wilcox, K. (1982). Ethnography as a methodology and its applications to the study of schooling: A review. In G. Spindler (Ed.), *Doing the ethnography of schooling* (pp. 456-458). New York: Holt, Rinehart, & Winston.

Witty, E. (1978). Equal educational opportunity for gifted minority group children: Promise or possibility? *Gifted Child Quarterly,* 22, 344-352.

Is the Proof in the Pudding?

Reanalyses of Torrance's (1958 to Present) Longitudinal Data

Jonathan A. Plucker, Indiana University

The use of divergent thinking (DT) tests to assess creativity has been strongly criticized in recent years. Several critics have noted that DT test scores have shown little evidence of predictive validity with respect to adult creative achievement. Data from Torrance's (1972a) elementary school longitudinal study (1958–present) were reanalyzed using structural equation modeling. Results suggest that just under half of the variance in adult creative achievement is explained by DT test scores, with the contribution of DT being more than three times that of intelligence quotients. However, comprehensive longitudinal models of creative achievement based on current creativity and cognitive theory have yet to be empirically validated.

Few areas of creativity research are as controversial or contentious as the validation of divergent thinking (DT) test scores. Current thought on the nature of creativity is gradually moving away from psychometric perspectives toward more postmodern approaches (e.g., Feldman, Csikszentmihalyi, & Gardner, 1994; Gardner, 1993). Furthermore, DT tests are criticized for a lack of discriminant and predictive validity (e.g., Gardner, 1988; Weisberg, 1993) and an overreliance on content generality (Baer, 1991, 1993). Taken collectively, the paradigm shift and mounting methodological criticisms have led many researchers and practitioners to avoid the use of DT measures (Plucker & Renzulli, 1999).

The decision to avoid studying and measuring DT may have been made too hastily. More than any other approach to the study of creativity, research on DT provides the foundation for creativity training programs in education and business (e.g., Basadur, Graen, & Green, 1982; Isaksen & Treffinger, 1985; Meeker & Meeker, 1982; Renzulli, 1976; Taylor, 1988). DT tests are also frequently used by educators to identify students with high creative potential (Hunsaker & Callahan, 1995; Runco,1993), and creativity researchers and theorists — both in the United States (Davis, 1989; Plucker & Renzulli, 1999; Runco, 1991, 1993) and across the globe (de Alencar, 1996; Kim & Michael, 1995; Niaz & De Nunez, 1991) — value DT and administer DT tests in the course of their work. Clearly, DT is still valued by a significant number of theorists, researchers, and practitioners. Rather than dismissing the psychometric study of creativity out of hand, researchers and practitioners would be better served by continuing thorough evaluations of DT tests' psychometric qualities (Plucker & Renzulli, 1999).

Reprinted from *Creativity Research Journal*, vol. 12, no. 2, 1999

Review of Literature

As early as 1964 in the scientific study of creativity, Taylor and Holland recommended longitudinal studies that "use a very wide variety of potential predictors, and then, after a suitable follow-up period, utilize good external criteria of creativity" (p. 48) in order to collect evidence of creativity measures' predictive validity. Yet a decade later, in an evaluation of efforts to establish DT tests' predictive validity, Wallach (1976) stated that "subjects vary widely and systematically in their attainments — yet little if any of that systematic variation is captured by individual differences on ideational fluency tests" (p. 60). Other assessments of the predictive validity of DT tests have been similarly negative and pessimistic (Baer, 1993, 1994; Gardner, 1988, 1993; Kogan & Pankove, 1974).

Researchers suggest several possible reasons for DT tests' apparent lack of predictive validity. The scores may be susceptible to various coaching and administration issues (Hattie, 1980; Wallach, 1976). Longitudinal studies may be too brief to allow people to achieve creatively (Torrance, 1972b, 1979). Studies may overemphasize quantity of creative achievement at the expense of quality of achievement (Runco, 1986). Statistical procedures may be inadequate for the analysis of complex longitudinal data (Hocevar & Bachelor, 1989; Plucker & Renzulli, 1999). Initial socioeconomic conditions and intervening life events (e.g., the "fourth grade slump") may make prediction of adult creative achievement primarily on the basis of ideational-thinking test scores difficult (Cramond, 1993, 1994; Torrance, 1981b). Another possible reason may be that the distributions of scores on creativity assessments are often nonnormally distributed, which violates the assumptions of many statistical procedures.

A majority of the possible reasons for weak predictive validity coefficients represent weaknesses in methodology more than weaknesses in the psychometric approach to creativity research. Indeed, researchers who have addressed at least a few of these weaknesses (e.g., Sawyers & Canestaro's 1989 domain-specific study) have collected positive evidence of predictive validity.

Case for Predictive Validity

The poor predictive power of DT tests is not universally accepted. To the contrary, several studies provide at least limited evidence of discriminant and pre-

---

I thank E. Paul Torrance for enthusiastically providing access to his data and discussing the original study, Ginette Delandshere for advice related to the statistical analyses, and Mark Runco for constructive criticisms related to the manuscript. The work and opinions expressed in the article are the responsibility of the author.

This article is based on a presentation made at the annual convention of the National Association for Gifted Children, November 9, 1997.

Correspondence and requests for reprints should be sent to Jonathan A. Plucker, Indiana University, School of Education, 201 North Rose Ave, Bloomington IN 47405-1006. E-mail: jplucker@indiana.edu.

dictive validity for DT tests (Howieson, 1981; Milgram & Hong, 1994; Milgram & Milgram, 1976; Okuda, Runco, & Berger, 1991; Rotter, Langland, & Berger, 1971; Runco, 1986; Torrance, 1969, 1972a, 1972b, 1981a, 1981b; Torrance & Safter, 1989; Torrance, Tan, & Allman, 1970; Torrance & Wu, 1981; Yamada & Tam, 1996), especially under certain sampling and assessment conditions (e.g., samples of high-IQ children that use content-specific DT measures; see Hocevar, 1981; Milgram & Milgram, 1976; Runco, 1986).

Where does the research leave the debate over the validity of DT tests? As noted, the research is rather equivocal. Additional research on the predictive validity of DT tests is needed, especially that addressing the weaknesses of previous longitudinal research. The purpose of this study was to address criticisms of DT test research through a reanalysis of data from a seminal longitudinal study of creativity.

Torrance Elementary School Study

E. Paul Torrance began his seminal research in response to skepticism surrounding the belief that creative children mature into creative adults. Although Torrance conducted several longitudinal studies of creativity (Torrance, 1969, 1972a, 1972b, 1993), his study of over 200 elementary-school students (Torrance, 1981b; Torrance & Wu, 1981) is among the most compelling. Data collection began in 1958, and the students are still being studied, with a new round of data collection currently underway (E. P. Torrance, personal communication, November 20, 1997). This investigation is one of the longest studies specifically focusing on creative abilities and achievement. As such, it is often cited in debates on the predictive validity of DT tests. Any analysis of this topic should begin with this seminal study.

The sample for the study initially included every student attending two Minnesota elementary schools from 1958-1965. Students completed the Torrance Tests of Creative Thinking (TTCT; Torrance & Ball, 1984) on an annual basis and provided additional demographic information. Intelligence and achievement test data were also gathered for a majority of the children. In 1980, 22 years after the initiation of the study, Torrance and his colleagues collected an additional set of data. This round of data collection was restricted to those students (*n* = 400) in the original sample who had completed the TTCT for three or more years (generally third through fifth grade).

Seventy percent of these individuals were located, and 220 (55 percent) participated in the second round of data collection. In 1980, the average age of this group was 27.5 years. Space limitations do not permit a more detailed description of the original study, but interested readers are referred to Cramond (1993) and Torrance (1981b).

Method

Sample

After removing cases with a significant amount of missing data, the sample for these reanalyses involved 212 people, 54.2 percent female. The average IQ score for the students, measured with the Stanford-Binet, Wechsler Intelligence Scale for Children, or California Test of Mental Maturity, was 121 (SD = 16).

Instrumentation

Although data collection in each round was exhaustive, a limited number of variables were available for reanalysis in this study. A small number of cases contained missing data for specific variables (roughly 10-15 percent of IQ scores and 1-5 percent of specific DT test scores). The mean for each variable was substituted for missing data when necessary.

Initial (1958-1965) data. Intelligence test scores were available for this reanalysis. IQ scores were introduced primarily as control variables, although the high percentage of missing school achievement data resulted in the exclusion of achievement scores from the reanalyses.

Each student completed the TTCT annually during grades 1 to 6. Three-year averages, which Torrance (1981b) considered to be the most stable estimate of DT ability, were used in the current investigation. Additionally, fifth-grade figural fluency, originality, and elaboration scores and verbal fluency, flexibility, and originality scores were available for most students. Unlike previous studies that utilized a total creativity index derived from the various scores (Torrance, 1981b; Yamada & Tam, 1996), each TTCT subtest score was included in the model, to form the latent predictor variables. Other potential predictor variables, such as future career image, foreign study, and involvement with a mentor, were excluded due to the high percentage of missing data (see Torrance, 1981b, 1987, for a description of these variables).

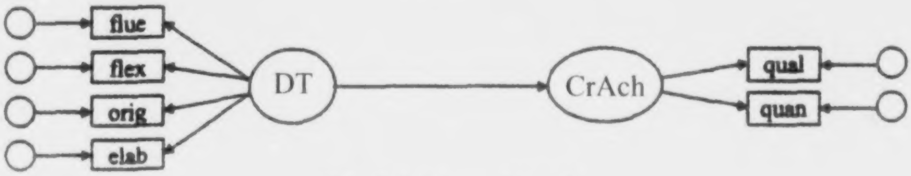
Follow-up creativity data. Two measures provided data for outcome variables in this reanalysis. The first was an estimate of the quantity of publicly recognized creative achievements such as inventions, published articles, awards for creativity, and other experiences similar to those found on other creative-activity checklists (Hocevar, 1979; Runco, 1987; Wallach & Wing, 1969). In addition, participants in the follow-up each provided a list of their three most creative achievements. Three judges rated these achievements on the basis of overall creativity in order to arrive at an estimate of creative achievement quality. These two measures provided the data that formed the latent outcome variables.

Data Analysis

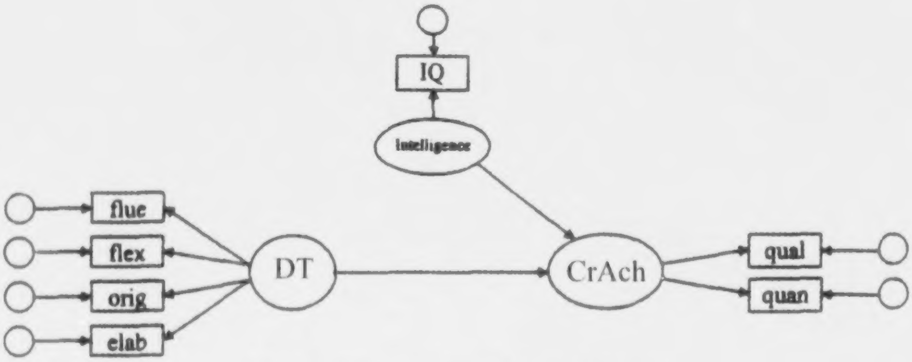
Most analyses of the data (e.g., Torrance, 1981b) have involved the use of bivariate correlations, and even more sophisticated attempts to analyze the data

(Yamada & Tam, 1996) have not controlled for students' intelligence. To capitalize on the use of latent variables, structural equation modeling was used to investigate the relation between DT test performance and subsequent creative achievement. The three general models used in this study appear in Figure 1. A second set of analyses, which included the fifth-grade figural form and verbal form scores, proceeded in the same manner as for the general models. Statistical bootstrapping was used to determine whether the nonnormal distributions of certain scores had an appreciable impact on the results. Covariance matrices for

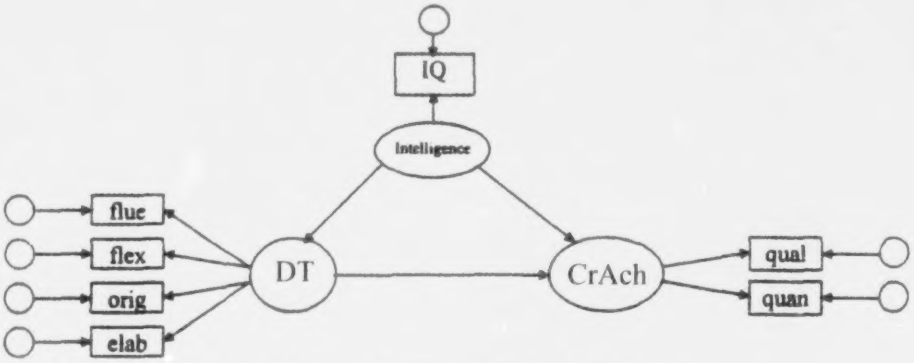
Figure 1



Model 1A — Simple



Model 1B — DT and Intelligence



Model 1C — Controlling for Intelligence

General divergent thinking (DT) models tests with structural equation modeling. CrAch = creative achievement; flue = fluency; flex = flexibility; orig = originality; elab = elaboration; qual = quality; quan = quantity.

Table 1

Goodness-of-Fit Indexes for the Four General Divergent Thinking Models

Model	x^2	df	x^2/df	NFI	TLI	RMSEA
Simple						
Model 1A	41.86	8	5.23	.93	.89	.14
Log(Quantity)	40.93	8	5.12	.94	.90	.14
Square Root(Quality)	41.84	8	5.23	.94	.90	.14
Creativity-Intelligence						
Model 1B	52.02	13	4.00	.91	.89	.12
Log(Quantity)	49.99	13	3.85	.93	.91	.12
Square Root(Quality)	51.52	13	3.96	.92	.90	.12
Complex						
Model 1C	44.23	12	3.69	.93	.90	.11
Log(Quantity)	42.19	12	3.52	.94	.92	.11
Square Root(Quality)	43.72	12	3.64	.94	.92	.11

Note: NFI = Normed Fit Index; TLI = Tucker-Lewis Index; RMSEA = root mean squared error of approximation.

the general DT and specific DT analyses are included in Appendices A and B, respectively.

Results

General DT Models

Three sets of models were fit to the data (Models 1A, 1B, and 1C in Figure 1). In Models 1B and 1C, the single measure of prior academic ability or achievement was an IQ score. This reliance on a single indicator factor to represent a complex construct such as intelligence is regrettable. Given the limitations of the dataset, however, the single indicator factor was unavoidable. To account for this weakness in Models 1B and 1C, the measurement error was fixed at .1 for the intelligence scores (implying a reliability estimate of .9). Although this technique is not a substitute for multiple indicators of intelligence, the reanalyses could proceed with reasonable faith in the results. Goodness-of-fit indexes for the general DT models appear in Table 1.

The simplest Model (1A), which posited a straightforward relation between the general DT and creative achievement latent variables, was included in the analyses for illustrative purposes (i.e., intelligence is not included in the model). Factor loadings for the latent variable indicators were moderate to large in magnitude, ranging from .54 to .79 for general DT and .68 to .72 for creative achievement. In the simple model, general DT accounted for 41 percent of the variance in creative achievement. Model 1A approached an acceptable goodness of fit (e.g., Tucker-Lewis Index > .90).

The next most complex Model (1B) included students' intelligence test scores as a second latent predictor variable. Again, the model was associated with marginally acceptable goodness-of-fit estimates. A similar amount of creative achievement variance was explained by Models 1A and 1B, with 40 percent explained in 1B. Interestingly, intelligence explained very little (3.6 percent) of this variance, and the path coefficient from general DT to creative achievement was considerably larger than the coefficient from intelligence to creative achievement (.60 vs. .19).

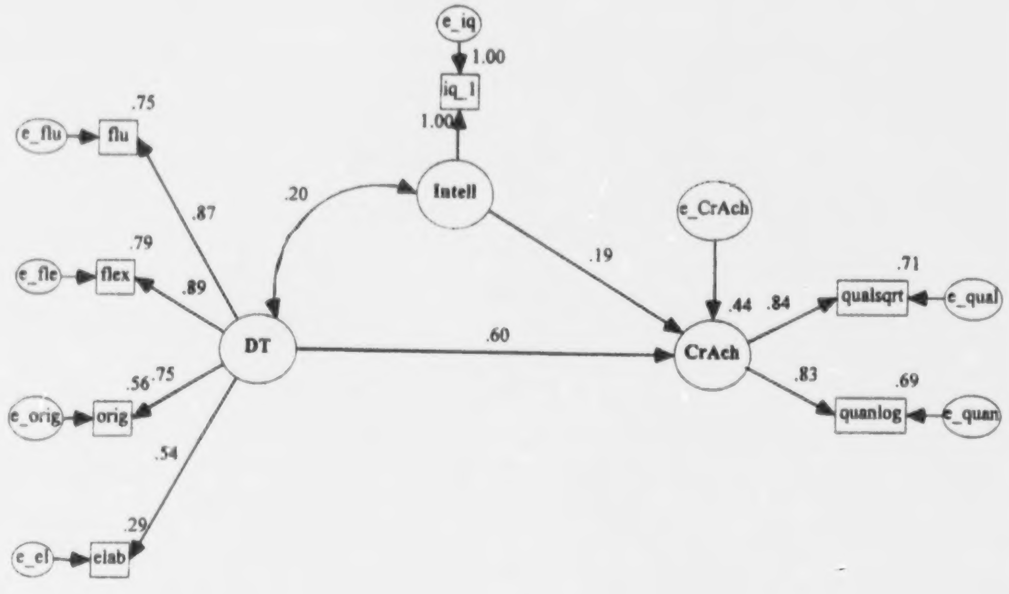
Several critics have argued that TTCT scores are correlated with intelligence test scores and are merely measuring rather limited aspects of intelligence. If this assumption is true, then TTCT scores should have little predictive power toward adult creative achievement when intelligence scores are controlled. To address this issue, the effect of prior intelligence was controlled in Model 1C by creating paths from the intelligence variable to general DT and creative achievement. Goodness-of-fit measures were relatively large. As a result, Model 1C was determined to have the superior fit of the tested models and was subjected to bootstrapping analysis.

Bootstrapping is a form of statistical resampling that can accomplish several purposes. By treating the study's sample as a population and drawing repeated samples from that population, model estimates can be recalculated and compared to the original "population" estimates. This resampling produces robust standard errors that can be applied to the original calculations, and the resampled and original estimates can be compared to determine the impact of nonnormal distributions. For example, if the original and resampled estimates differ significantly, then the distribution of one or more variable scores is probably nonnormal and should be transformed to remove this bias.

The bootstrapping procedure produced little evidence of bias for most of the variable scores. In fact, bias was generally very small: less than or approximately equal to the applicable standard error for all but the quantity and quality scores. Analysis of both relevant histograms revealed that the quantity distribution was slightly skewed and the quality distribution was kurtotic. To address this departure from normality, quantity scores were transformed logarithmically and the square root was taken for the quality scores.

Each of the three models (1A, 1B, and 1C) was refit to the data with the quantity transformation, represented by log(quantity) in Table 1, and both transformations, represented by sqrt(quality). In all three cases, the model with the quantity transformation had the best fit to the data, although the differences within each group of models were practically insignificant. Among the three groups of models, Model 1C had the best fit for the original data and both trans-

Figure 2



General divergent thinking (DT) Model 1C with transformed quantity and quality scores. Intell = intelligence; CrAch = creative achievement; qualsqrt = square root transformation of creative achievement quality scores: quanlog = logarithmic transformation of creative achievement quantity scores; flu =fluency; flex=flexibility; orig =originality; elab = elaboration. x2 = 43.72, df=12, p=.00. Bentler-Bonett Index = .93: Tucker-Lewis Index = .91.

formed sets of data. Figure 2 contains the standardized solution for Model 1C (with the transformed quantity and quality scores). Similar to the earlier models, just under half of the variance in adult creative achievement is explained by this model, with the contribution of general DT more than three times that of Intelligence. Analysis of standard errors and critical ratios suggests that the parameter estimates are significant at $p < .01$.

Specific Models

Analyses of the TTCT verbal and figural data proceeded along similar lines. Given the results of the general DT analyses, the transformed scores representing quantity and quality of creative achievement were used throughout the specific DT analyses. Table 2 contains the various goodness-of-fit measures used to compare the four sets of models.

The simple Model (2A) is similar to general DT Model 1B, with three uncorrelated latent variables (verbal DT, figural DT, and intelligence) predicting creative achievement. The Intelligence Controlled Model (2B) was created by adding paths from the intelligence latent variable to the other three latent variables, effectively controlling for intelligence. The third Model (2C) has the three correlated latent variables predicting creative achievement.

Table 2

Goodness-of-Fit Indexes for the Three Specific Divergent Thinking Models

Model	χ^2	df	χ^2/df	NFI	TLI	RMSEA
Simple Model (2A)	69.80	25	2.79	.90	.90	.09
Controlling for Intelligence (2B)	66.22	23	2.88	.90	.90	.09
Correlated (2C)						
Three indicators	47.31	22	2.15	.93	.94	.07
Two indicators	30.81	17	1.81	.95	.96	.06
Correlated, CU (2D)	46.43	21	2.21	.93	.93	.08

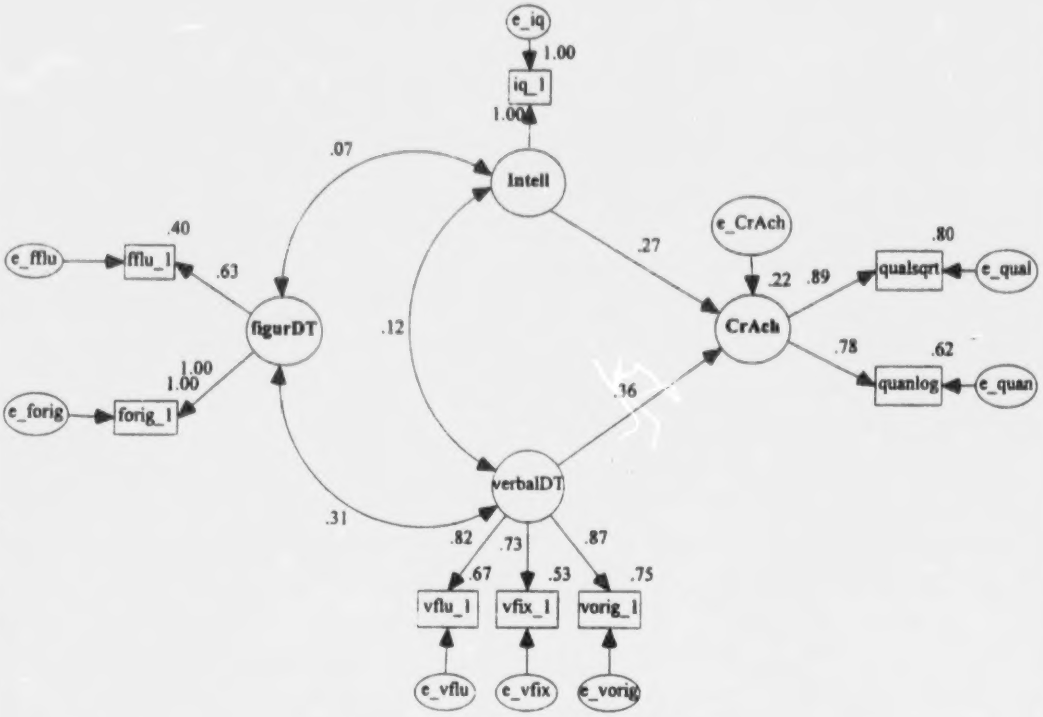
Note: In all specific divergent thinking models, creative achievement quality scores were subjected to a square root transformation, and creative achievement quantity scores were subjected to a logarithmic transformation. NFI = Normed Ft Index; TLI = Tucker-Lewis Index; RMSEA = root mean squared error of approximation; CU = correlated uniquenesses.

The fourth Model (2D) posits the presence of correlated uniquenesses (CU). The CU model hypothesizes that error for similar observed constructs is related. In this study, uniquenesses (i.e., error terms) for verbal and figural originality scores are correlated, as are uniquenesses for verbal and figural fluency scores. Additional models produced negative variances and other untenable parameter estimates and were, therefore, not statistically viable. Interestingly, these models included those derived from suggestions made by Hocevar and Bachelor (1989) in which latent variables representing fluency and originality were used rather than latent figural and verbal DT variables. Future research should include investigations of the appropriateness of these models.

In general, models were characterized by (a) evidence of low reliability for the figural elaboration scores, (b) low correlations between intelligence and the DT latent variables, (c) a moderate correlation between specific and verbal DT, (d) a low path coefficient between specific DT and creative achievement, and (e) a larger path coefficient between verbal DT and creative achievement than between intelligence and creative achievement. The correlated predictor variable Model (2C) had the superior fit of the tested models. The correlations between uniquenesses in Model 2D were very low, resulting in a relatively poor fit in comparison to Model 2C.

In response to the low reliability estimates for the figural elaboration scores, Model 2C was modified to include only fluency and originality as indicators of specific DT. Although omitting the elaboration scores required the specific DT-creative achievement path to be fixed at 0, the fit of the resulting model was considerably better than the other specific DT models. Figure 3 contains structural equation modeling parameter estimates for this correlated latent predictor variable, two-indicator specific DT model. Standard errors and critical ratios indicated that all parameter estimates were statistically significant at $p < .01$, with the exception of the correlations between intelligence and each DT latent variable. Bootstrapping of this model provided evidence of robust standard errors and very little bias between the original and bootstrapped parameter estimates.

Figure 3



Specific divergent thinking (DT) Model 2C with two figural DT indicator variables. Intell = intelligence; CrAch = creative achievement; qualsqrt = square root transformation of creative achievement quality scores; quanlog = logarithmic transformation of creative achievement quantity scores: verbalDT = verbal divergent thinking; vflu_1 = verbal fluency; vfix_1 = verbal flexibility; vorig-1 = verbal originality; figurDT = figural divergent thinking; fflu_1 = figural fluency; forig_1 = figural originality. XI = 30.81, df =17, p = .02. Bentler-Bonett Index = 95; Tucker-Lewis Index = .96.

Discussion

Every study has limitations, and any reanalysis of data is hardly an exception to that rule. Collectively, the students in the sample had higher than average intelligence, and only a portion of the original sample from the 1950s participated in the follow-up study. However, longitudinal studies are invariably marked by sampling issues, and lengthy studies (e.g., Terman's (1926) seminal research with high-IQ children; this study's duration of several decades) will probably have less than perfect sampling. Rather than be discarded, the results should be

BEST COPY AVAILABLE

evaluated with the unique characteristics of the sample in mind.

Rigorous statistical reanalyses of the longitudinal Torrance data provide evidence that directly refutes Wallach's (1976) assertion that "little if any of that systematic variation [in adult creative achievement] is captured by individual differences on ideational fluency tests" (p. 60). Indeed, the results, specifically the strong predictive power of TTCT scores relative to IQ estimates, support Torrance's (1981b) original conclusions about the predictive validity of DT tests.

The creativity-intelligence relationship has been of special interest to psychologists and educators. In an earlier reanalysis of the Torrance data, Yamada and Tam (1996) concluded that "higher intelligence is necessary for publicly recognized creative achievements" (p. 147). The present reanalysis suggests that intelligence may be a component of creative achievement, but it is a relatively weak one when measures of DT are considered. Granted, students in this study are generally quite intelligent, and additional research with a sample of more representative intelligence is necessary before any conclusions can be drawn regarding the relative importance of intelligence and DT to future creative achievement.

The results regarding figural and verbal DT are much more difficult to interpret. Although verbal DT was a better predictor of creative achievement than intelligence, figural DT was not a factor in the model. In sharp contrast to the general DT model, little of the creative achievement variance was explained by the intelligence and DT latent variables.

Three possible explanations for the anomalous findings come to mind. First, as suggested by Torrance and Ball (1984), the 3-year averages that provided indicators for general DT in the first set of analyses may indeed be more stable than one-shot administrations. Given the possible existence of the fourth-grade slump in creativity (Torrance, 1968), the restriction of specific DT scores to fifth-grade TTCT scores may have introduced considerable bias into the second set of analyses.

Second, the importance of verbal DT relative to figural DT may be due to a linguistic bias in the adult creative achievement checklists. For example, if a majority of the creative achievements required a high degree of linguistic talent, as opposed to spatial talent or problem solving talents, the verbal DT tests would be expected to have a significantly higher correlation to these types of achievement that other forms of DT. Of course, this same explanation could be used to argue that many intelligence tests are suspected to have a linguistic bias, so the intelligence scores should be expected to have a higher correlation to creative achievement than was found in this study. Research that utilizes multiple stable indicators of DT and intelligence would help elucidate some of these issues.

Third, DT and intelligence may simply not be highly or even moderately correlated. Wallach and Wing (1969) and Runco and Albert (1986) found that intelligence estimates and DT scores were not highly correlated, and the results of this study can be viewed as further evidence that the relationship between IQ and DT is negligible. All three possible explanations probably have some merit, and future research efforts should attempt to replicate these seemingly anomalous results.

Although the results of this study support the predictive validity of DT tests, the conclusion that the TTCT is "the best predictor for adult creative achievement" (Yamada & Tam, 1996, p. 147) is a bit premature. Very few longitudinal studies involving the TTCT or other DT tests have considered the impact of personality variables as predictors of creativity. The models tested in this study would almost certainly have been improved (i.e., they would have explained a greater percentage of the variance of future creative achievement) if personality and socioeconomic vanables were considered.

However, as Renzulli (1991) noted, educators have little control over the personality characteristics of children. As a result, teachers' major concerns are divergent and convergent thinking, problem finding and problem solving, and other skills that can be directly enhanced. In short, educators can help enhance a child's DT abilities, and even raise his or her creative self-efficacy (see Amabile, 1983; Hennessey & Amabile, 1987), but we have little control over a student's personality and demographic characteristics. An alternative to the Yamada and Tam (1996) conclusion would be to state that DT tests appear to be the best cognitive predictor of creative achievement over which we can have an appreciable educational impact. This more conservative assessment appears to be justified at this time.

Designing Future Predictive Validity Research

The educational implications of including personality variables in future longitudinal studies may be few and far between, but the research implications are substantial. Given the proliferation of creativity theory and models in the past 10 years, one would expect these models to have been tested empirically. However, this is rarely the case. For example, Amabile (1983) proposed a very influential model for the social psychology of creativity, but this model has yet to be tested longitudinally.

With this in mind, future longitudinal studies should consider personality and demographic characteristics as they influence DT and other creative abilities and attitudes, but not as means of and by themselves. Useful models such as Marsh's (1986, 1990) internal-external frame of reference model of self-concept should also be considered when researchers attempt to improve upon efforts to predict creative achievement.

Gardner (1988) noted a reliance on self-reported data as outcome measures in longitudinal creativity research. Although researchers have addressed this perceived weakness in psychometric studies with little apparent change in their conclusions (Runco, 1986; Torrance & Ball, 1984), researchers involved with longitudinal research should address Gardner's criticism about the use of psychometric outcomes in psychometric studies. Given Gardner's criticism and recent research that suggests that the methods used to measure creativity predetermine some of its observed characteristics (Plucker, 1998, 1999), inclusion of multiple methodologies for collecting creativity data is a wise decision (Hocevar & Bachelor, 1989). For example, rather than rely solely on the administration of the TTCT, researchers could administer a battery of DT tests, including the Wallach and Kogan (1965) exercises. Additionally, inclusion of data collected with the consensual assessment technique (Amabile, 1983) and related performance-based assessments of creativity would strengthen the design of a longitudinal creativity study. The use of multiple indicators is especially useful when advanced statistical techniques, such as structural equation modeling, are used to analyze longitudinal data. Furthermore, other criteria (e.g., attitudes toward various aspects of ideation and creativity) may be more appropriate to include as outcomes in longitudinal models.

Conclusions

DT tests have recently fallen out of favor, especially in educational settings. In particular, critics have pointed out that the Torrance Tests of Creative Thinking (Torrance & Ball, 1984; and the longitudinal studies that Torrance used to establish evidence of their validity) are fraught with methodological weaknesses. The results of this reanalysis, even considering its limitations, are relatively supportive of Torrance's original conclusions regarding the ability of DT test scores to predict creative achievement. In fact the 3-year averages of general DT scores were considerably better predictors of adult creative achievement than intelligence test scores. Furthermore, latent variables representing DT were not correlated highly with intelligence, further supporting the claims of DT proponents that DT and intelligence represent relatively independent constructs — at least in high-IQ samples.

Considering the results of this study and other recent research, assessments of DT appear to be associated with acceptable levels of predictive validity. Indeed, if creative achievement is the focus of an educational or empirical effort, DT tests appear to be preferred as a measure to intelligence tests. Of course, DT skeptics are often critical of traditional methods for measuring intelligence, which is not surprising, given the critics' predominantly postmodern perspective. The issue at hand may involve the appropriateness of psychological measurement as opposed to the assessment of DT.

This is not to say that critics of DT measures do not have many viable points. DT research has only recently included multiple measures of creativity (e.g., the Consensual Assessment Technique, Amabile, 1983; multiple types of DT measures); statistical design has often been lacking in quality, and psychological measures can always be improved. Indeed, many of the leading DT researchers have invested significant efforts in the improvement of DT tests and their administration (Runco & Mraz, 1992; Runco & Okuda, 1991; Runco, Okuda, & Thurston, 1987; Torrance, 1988). Longitudinal studies that are initiated today will be more methodologically sound than past research, due to improvements in research design and data analysis and to the work of DT researchers. Studies such as Torrance's exhaustive work and reanalyses such as the present one provide additional suggestions for future researchers: Use multiple indicators of each targeted construct; test models that included demographic, environmental, and social variables; and attempt to minimize the amount of missing data so that powerful techniques such as bootstrapping can be used to analyze the data.

References

Amabile, T. M. (1983). *The social psychology of creativity.* New York: Springer-Verlag.

Baer, J. (1991). Generality of creativity across performance domains. *Creativity Research Journal, 4, 23-39.*

Baer, J. (1993. December/January). Why you shouldn't trust creativity tests. *Educational Leadership, 51(4), 80-83.*

Baer, J. (1994, October). Why you still shouldn't trust creativity tests. *Educational Leadership, 52(2), 72-73.*

Basadur, M. S., Graen, G. B., & Green. S. G. (1982). Training in creative problem solving: Effects on ideation and problem finding in an applied research organization. *Organizational Behavior and Human Performance, 30, 41-70.*

Cramond, B. (1993). The Torrance Tests of Creative Thinking: From design through establishment of predictive validity. In R F. Subotnik & K. D. Arnold (Eds.). *Beyond Terman: Contemporary longitudinal studies of giftedness and talent* (pp. 229-254). Norwood, NJ: Ablex.

Cramond, B. (1994. October). We *can* trust creativity tests. *Educational leadership, 52(2), 70-71.*

Davis, G. A. (1989). Testing for creative potential. *Contemporary Educational Psychology, 14, 257-274.*

de Alencar, E. M. L. S. (1996). University students' evaluation of their own creativity and their teachers' and colleagues' level of creativity. *Gifted Education International, 11, 128-130.*

Feldman, D. H.. Csikszentmihalyi, M., & Gardner, H. (1994). *Changing the world: A framework for the study of creativity.* Westport, CT: Praeger.

Gardner. H. (1988). Creativity: An interdisciplinary perspective. *Creativity Research Journal, 1, 8-26.*

Gardner, H. (1993). *Creating minds.* New York: Basic Books.

Hattie, J. (1980). Should creativity tests be administered under testlike conditions? An empirical study of three alternative conditions. *Journal of Educational Psychology, 72. 87-98.*

Hennessey, B. A., & Amabile, T. M. (1987). *Creativity and learning.* Washington, DC: National Education Association.

Hocevar, D. (1979, April). *The development of the Creative Behavior Inventory.* Paper presented at the annual meeting of the Rocky Mountain Psychological Association (ERIC Document Reproduction Service No. ED 170 350).

Hocevar, D. (1981). Measurement of creativity: Review and critique. *Journal of Personality Assessment, 45, 450-464.*

Hocevar, D., & Bachelor, *P.* (1989). A taxonomy and critique of measurements used in the study of creativity. In J. A. Glover, R R. Ronning, & C. R. Reynolds (Eds.), *Handbook of creativity* (pp. 53-75). New York: Plenum.

Howieson, N. (1981). A longitudinal study of creativity — 1965-1975. *Journal of Creative Behavior, 75, 117-134.*

Hunsaker, S. L., & Callahan, C. M. (1995). Creativity and giftedness: Published instrument uses and abuses. *Gifted Child Quarterly, 39, 110-114.*

Isaksen, S. G., & Treffinger, D. J. (1985). *Creative problem-solving: The basic course.* Buffalo, NY: Bearly

Limited.

Kim, J., & Michael, W. B. (1995). The relationship of creativity measures to school achievement and to preferred learning and thinking style in a sample of Korean high school students. *Educational and Psychological Measurement, 55,* 6074.

Kogan, N., & Pankove, E. (1974). Long-term predictive validity of divergent-thinking tests: Some negative evidence. *Journal of Educational Psychology, 66,* 802-810.

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal, 23,* 129-149.

Marsh, H. W. (1990). Influences of internal and external frames of reference on the formation of math and English self-concepts. *Journal of Educational Psychology, 82,* 107-116.

Meeker, M., & Meeker, R (1982). *Structure-of-Intellect Learning Abilities Test: Evaluation, leadership, and creative thinking.* El Segundo, CA: SOI Institute.

Milgram. R M., & Hong, E. (1994). Creative thinking and creative performance in adolescents as predictors of creative attainments in adults: A follow-up study after 18 years. In R. F. Subotnik & K. D. Arnold (Eds.), *Beyond Terman: Contemporary longitudinal studies of giftedness and talent* (pp. 212-228). Norwood, NJ: Ablex.

Milgram, R. M., & Milgram, N. A. (1976). Creative thinking and creative performance in Israeli students. *Journal of Educational Psychology, 68,* 255-259.

Niaz, M., & De Nunez, G. S. (1991). The relationship of mobility-fixity to creativity, formal reasoning and intelligence. *Journal of Creative Behavior, 25,* 205-217.

Okuda, S. M., Runco, M. A., & Berger, D. E. (1991). Creativity and the finding and solving of real-world problems. *Journal of Psychoeducational Assessment, 9,* 45-53.

Plucker, J. A. (1998). Beware of simple conclusions: The case for content generality of creativity. *Creativity Research Journal,* 11, 179-182.

Plucker, J. (1999). Reanalyses of student responses to creativity checklists: Evidence of content generality. *Journal of Creative Behavior, 33,* 126-137.

Plucker, J., & Renzulli, J. S. (1999). Psychometric approaches to the study of creativity. In R. J. Sternberg (Ed.), *Handbook of human creativity* (pp. 35-60). New York: Cambridge University Press.

Renzulli, J. S. (1976). *New directions in creativity.* New York: Harper & Row.

Renzulli, J. S. (1991, August). *A general theory for the development of creative productivity through the pursuit of ideal acts of learning.* Paper presented at the biennial meeting of the World Congress for the Gifted and Talented, The Hague, Netherlands.

Rotter, D. M., Langland, L., & Berger, D. (1971). The validity of tests of creative thinking in 7-year-old children. *Gifted Child Quarterly, 4,* 273-278.

Runco, M. A. (1986). Divergent thinking and creative performance in gifted and nongifted children. *Educational and Psychological Measurement, 46,* 375-384.

Runco, M. A. (1987). The generality of creative performance in gifted and nongifted children. *Gifted Child Quarterly, 31,* 121125.

Runco, M. A. (1991). Comment on investment and economic theories of creativity: A reply to Sternberg and Lubart. *Creativity Research Journal, 4,* 202-205.

Runco, M. A. (1993). Divergent thinking, creativity, and giftedness. *Gifted Child Quarterly, 37,* 16-22.

Runco, M. A., & Albert, R. S. (1986). The threshold theory regarding creativity and intelligence: An empirical test with gifted and nongifted children. *The Creative Child and Adult Quarterly,* 11, 212-218.

Runco, M. A., & Mraz, W. (1992). Scoring divergent thinking tests using total ideational output and a creativity index. *Educational and Psychological Measurement, 52,* 213-221.

Runco, M. A., & Okuda, S. M. (1991). The instructional enhancement of the flexibility and originality scores of divergent thinking tests. *Applied Cognitive Psychology, 5,* 435-441.

Runco, M. A., Okuda, S. M., & Thurston, B. J. (1987). The psychometric properties of four systems for scoring divergent thinking tests. *Journal of Psychoeducational Assessment, 2,* 149-156.

Sawyers, J. K., & Canestaro, N. C. (1989). Creativity and achievement in design coursework. *Creativity Research Journal, 2,* 126-133.

Taylor. C. W. (1988). Various approaoaches to and definitions of creativity. In R J. Sternberg (Ed.), *The nature of creativity: Contemporary psychological perspectives* (pp. 99-121). New York: Cambridge University Press.

Taylor, C. W., & Holland, J. (1964). Predictors of creative performance. In C. W. Taylor (Ed.), *Creativity: Progress and potential* (pp. 15-48). New York: McGraw-Hill.

BEST COPY AVAILABLE

Terman, L. M. (1926). *Mental and physical traits of a thousand gifted children* (2nd ed.). Stanford, CA: Stanford University Press.

Torrance, E. P. (1968). A longitudinal examination of the fourth grade slump in creativity. *Gifted Child Quarterly, 12,* 195-199.

Torrance, E. P. (1969). Prediction of adult creative achievement among high school seniors. *Gifted Child Quarterly, 13,* 223-229.

Torrance, E. P. (1972a). Career patterns and peak creative achievements of creative high school students 12 years later. *Gifted Child Quarterly, 16,* 75-88.

Torrance, E. P. (1972b). Predictive validity of the Torrance Tests of Creative Thinking. *Journal of Creative Behavior, 6,* 236-252.

Torrance, E. P. (1979). Unique needs of the creative child and adult. In A. H. Passow (Ed.), *The gifted and talented: Their education and development. 78th NSSE Yearbook* (pp. 352-371). Chicago: The National Society for the Study of Education.

Torrance, E. P. (1981a). Empirical validation of criterion-referenced indicators of creative ability through a longitudinal study. *Creative Child and Adult Quarterly, 6,* 136-140.

Torrance, E. P. (1981b). Predicting the creativity of elementary school children (1958-1980) — and the teacher who "made a difference." *Gifted Child Quarterly, 25,* 55-62.

Torrance, E. P. (1987). Future career image as a predictor of creative achievement in a 22-year longitudinal study. *Psychological Reports, 60,* 574.

Torrance, E. P. (1988). The nature of creativity as manifest in its testing. In R. J. Sternberg (Ed.), *The nature of creativity: Contemporary psychological perspectives* (pp. 43-75). New York: Cambridge University Press.

Torrance, E. P. (1993). The beyonders in a 30 year longitudinal study of creative achievement. *Roeper Review, 15,* 131-135.

Torrance, E. P., & Ball, O. E. (1984). *Torrance Tests* of *Creative Thinking: Revised manual.* Bensenville, IL: Scholastic Testing Services.

Torrance, E. P., & Safter, H. T. (1989). The long range predictive validity of the Just Suppose Test. *Journal of Creative Behavior, 23,* 219-223.

Torrance. E. P., Tan, C. A., & Allman, T. (1970). Verbal originality and teacher behavior: A predictive validity study. *Journal of Teacher Education, 21,* 335-341.

Torrance, E. P., & Wu, T. H. (1981). A comparative longitudinal study of the adult creative achievement of elementary school children identified as highly intelligent and as highly creative. *Creative Child and Adult Quarterly, 6,* 71-76.

Wallach, M. A. (1976, January-February). Tests tell us little about talent. *American Scientist,* 57-63.

Wallach, M A., & Kogan, N. (1965). *Modes of thinking in young children: A study of the creativity-intelligence distinction.* New York: Holt, Rinehart & Winston.

Wallach, M. A., & Wing, C. W., Jr. (1969). *The talented student: A validation of the creativity-intelligence distinction.* New York: Holt, Rinehart & Winston.

Weisberg, R W. (1993). *Creativity: Beyond the myth of genius.* New York: Freeman.

Yamada, H., & Tam, A. Y. W. (1996). Prediction study of adult creative achievement: Torrance's longitudinal study of creativity revisited. *Journal of Creative Behavior, 30,* 144-149.

Appendix A
Covariance Matrix Used in General Divergent Thinking Analysis

	IQ	Quantity	Quality	Fluency	Flexibility	Originality	Elaboration
IQ	195.244						
Quantity	1.769	0.282					
Quality	3.287	0.311	0.703				
Fluency	17.916	1.788	2.521	54.622			
Flexibility	19.346	1.822	2.889	43.524	55.467		
Originality	16.546	1.716	3.067	38.020	39.626	66.862	
Elaboration	10.741	1.985	3.046	27.095	27.143	39.572	76.022
M	121	0.849	3.823	50.316	50.401	48.967	48.042
SD	16	0.532	0.840	7.408	7.465	8.196	8.740

Note: Quantity = log(quantity); Quality = square root (quality).

Appendix B

Covariance Matrix Used in Specific Divergent Thinking (DT) Analyses

	Verbal						Figural		
	Originality	**Flexibility**	**Fluency**	**IQ**	**Quantity**	**Quality**	**Fluency**	**Originality**	**Elaboration**
Verbal DT									
Originality	634.074								
Flexibility	213.023	183.069							
Fluency	408.606	189.094	528.357						
IQ	47.198	10.603	27.660	195.244					
Quantity	4.070	1.667	2.459	1.769	0.282				
Quality	6.768	3.668	4.157	3.287	0.311	0.703			
Figural									
Fluency	35.178	0.635	37.865	-4.271	0.266	0.051	50.607		
Originality	36.668	7.228	38.931	5.106	0.384	0.373	23.875	28.097	
Elaboration	18.317	4.720	19.045	2.143	0.297	0.495	6.054	6.370	6.735
M	37.384	23.113	51.339	121	0.849	3.823	21.720	14.912	9.401
SD	25.240	13.562	23.040	16	0.532	0.840	7.131	5.313	2.601

Note: Quantity = log (quantity); Quality = square root (quality).

Rorschach Interpretation With High-Ability Adolescent Females: Psychopathology or Creative Thinking?

Kristin W. Franklin and Dewey G. Cornell,
University of Virginia

Highly intelligent and creative persons have long posed interpretation difficulties for users of the Rorschach Inkblot Test. This study examined Exner's (1993) Schizophrenia, Depression, and Coping Deficit indices as adjustment measures in a sample of 43 female adolescents enrolled in an early college entrance program and a comparison group of 19 girls enrolled in public high school gifted programs. Contrary to conventional interpretation, higher scores on the Rorschach Schizophrenia Index among the accelerants were correlated with healthy emotional adjustment on both the California Psychological Inventory and the Self-Perception Profile for Adolescents (SPPA). Further analyses offered support for the hypothesis that among accelerants, elevated scores on the Rorschach constellations did not indicate psychopathology, but rather their creative thinking style.

How do we interpret the Rorschach responses of unusually creative and intelligent persons? Rorschach researchers (Barrett, 1957; Gallagher & Crowder, 1957; Gallucci, 1989) have long observed that particularly creative, intelligent individuals can produce strikingly original, often quite elaborate and unorthodox Rorschach responses that must be carefully distinguished from the responses of emotionally disturbed or psychotic individuals. The distinction between a richly creative response and a psychopathological one may rest not in a single characteristic or score, but in a combination or constellation of factors.

Three of Exner's (1993) special indices — the Schizophrenia Index, the Depression Index, and the Coping Deficit Index — represent especially ambitious efforts to combine a series of scores and response characteristics into more discriminating indicators of psychopathology. There is substantial evidence that these indices accurately discriminate between patient and nonpatient samples, but the performance of a sample selected for high intelligence and creativity has not been considered. Moreover, the indices were developed on adult samples, and only a few studies have applied them to adolescents. This study investigated the Rorschach performance of female early college entrants, a group that produced an unusually rich series of protocols while participating in a study of the effects of early college entrance on personality adjustment.

Reprinted from the *Journal of Personality Assessment,* 1997, *68*(1)

Special Indices

The Schizophrenia Index was developed on the assumption that schizophrenic individuals would show impairment in four areas — inaccurate perception, disordered thinking, inadequate controls, and interpersonal ineptness — using Rorschach variables including measures of form quality, human movement, and Special Scores (Exner, 1993). The index has been tested with samples of schizophrenics and nonschizophrenics, and accuracy rates fell within the range of 72 percent to 89 percent, with false positive rates ranging from zero to 11 percent (Exner, 1993). None of the 585 participants in Exner's normative samples of adolescent nonpatients show elevated Schizophrenia scores. In Netter and Viglione's (1994) study, the Schizophrenia Index correctly identified 14 of 20 schizophrenics, while misidentifying only three of 20 controls. However, in a study of adolescent inpatients, Archer and Gordon (1988) correctly diagnosed only 47 percent of schizophrenics using Exner's index. Additionally, 31 of 82 nonschizophrenics received Schizophrenia scores of 4 or more.

The Depression Index was created by identifying variables that discriminated depressed persons from nonpatients. The index was effective in distinguishing depressed persons, with 81 percent of Exner's depressed sample showing elevated values, and false positive rates in the 2-3 percent range (Exner, 1993). Less than 1 percent of Exner's adolescent norm group (Exner, 1990) produced Depression scores in the elevated range. However, other studies have shown the Depression Index to be less effective. The Depression Index correctly identified only seven of 67 depressed patients in Archer and Gordon's (1988) study. However, their rate of false positives was low as well, with only five of 121 nondepressed patients receiving elevated Depression scores.

The Coping Deficit Index (CDI) was developed in an effort to improve on the Depression Index. Individuals who score high on this index are likely to have impoverished or unrewarding social relationships, and have difficulty coping with social demands (Exner,1993). The index correctly identified 81 percent of a patient group defined as "helpless in the face of contending with a complex society" (Exner,1993, p. 362); with a false positive rate among nonpatient adults of 3 percent, and among nonpatient children, 6-24 percent. Less than 1 percent of Exner's adolescent norm group showed elevated CDI scores.

Rorschach Performance of High-Ability Subjects

Research on the Rorschach performance of high-ability subjects has produced mixed results. Early studies (Gallagher & Crowder, 1957; Selig, 1958) reported a high incidence of disturbed thinking among samples selected for high intelligence. More recently, Gallucci (1989) investigated 72 adolescents with IQs higher than 135 and found a preponderance of odd or strange responses *(DV* and *DR),* as well as significantly more unusual details (*Dd*) than a control group of average IQ adolescents. His high-ability adolescents also scored higher on the X-% index, and 72 percent of the sample were identified as positive for schizophrenia on an earlier version of the Exner Schizophrenia Index. Although their Rorschach results were indicative of psychopathology, the high IQ adolescents did not differ from the control group in behavior problems as measured by the Child Behavior Checklist (CBC). Furthermore, Rorschach scores presumed to be indicative of psychopathology did not correlate with any of the CBC scales. Gallucci cautioned against the conclusion that the unusual Rorschach performance of high-ability adolescents truly reflects psychopathology.

One possible explanation for the seemingly pathological Rorschach performance of high-ability subjects may lie in their creative thinking style. According to Sternberg and Davidson (1983), high-ability individuals often encode and process information in a creative, autonomous, and therefore unusual fashion. Rorschach responses that reflect a creative thinking style might be interpreted as pathological, because unusual and divergent responses are often scored as less reality-based, deviant, and indicative of schizophrenia or thought disorder.

Kris (1952) theorized that creative individuals have a capacity to relax normal ego defenses and use primary process in a productive manner, a process described as "regression in the service of the ego." Holt (1977) applied Kris's formulation to Rorschach performance through the development of scales to measure adaptive regression. Adaptive regression refers to the degree to which primary process responses, which customarily either contain drive-laden, aggressive, or sexual content, or involve illogical thinking processes, are integrated in an adaptive fashion (Dudek & Chamberland-Bouhadana, 1982). Holt (1970) developed a scoring system for adaptive regression, operationally defined to consider both the amount of primary process in a response and the adaptive defensiveness evident in the response.

Although mentally disordered subjects exhibit a high level of primary process on the Rorschach (Dudek, 1970), creative subjects have been found to integrate primary process responses adaptively (Pine & Holt, 1960; Dudek & Chamberland-Bouhadana, 1982). Rorschach responses reflecting adaptive regression meet two criteria: first, they are either aggressive or sexual in nature, or are of deviant form; and second, they are then defended in a realistic, socially acceptable manner. For example, a response containing aggressive primary process such as "a cannibal feast" can be logically defended and made more acceptable by referring to its cultural context: "a cannibal feast… they look like African natives" (Holt, 1970, p. 177).

This study concerns the Rorschach performance of a group of adolescent girls who entered college at ages 13 to 17. Girls admitted to the early college entrance program were selected not only on the basis of high intelligence and academic aptitude, but the presence of other personal characteristics judged to be important in undertaking a nonconventional, accelerated education, including independence, maturity, and creativity. Previous researchers have noted the non-conformist and independent styles of early college entrants (Cornell, Callahan, & Loyd, 1991b; Robinson & Janos, 1986).

There are conflicting findings regarding the social and emotional adjustment of high-ability adolescents who enter college at an early age. Several studies have found no differences between accelerants and nonaccelerants on a variety of adjustment measures (Brody & Benbow, 1987; Janos et al., 1988; Robinson & Janos, 1986). Using the California Psychological Inventory (CPI), Cornell et al. (1991b) found early college entrants to be independent, resourceful, and self-assured.

Other authors have found evidence of adjustment difficulties, low self-esteem, and friendship difficulties among early college entrants. Cornell, Callahan, and Loyd (1991a) found evidence of socioemotional adjustment problems, including depression and suicidality, among some early college entrants. Lubkowski, Whitmore, and Ramsay (1992) observed a drop in self-esteem after the first semester of the accelerants' college education. The authors noted that the students' drop in self-esteem might be due to the fact that academic achievements might not have come to them as easily as they had previously. In addition, when comparing themselves with other advanced students, early college entrants' image of themselves in the academic arena may diminish. There has also been concern that early college entrants' social adjustment may suffer. Janos et al. (1988) found that early college entrants in their first and second years of college prefer to interact with other early college entrants, and had limited friendships with regular college students. However, after the third year, the accelerants expanded their friendships to include their older classmates.

Accelerants were administered the Rorschach during their first month of college as part of a longitudinal study of the psychological adjustment of early college entrants. Previous reports have documented the incidence of emotional problems among some of the students (Cornell et al., 1991b), as well as described the overall growth and improvement in personality adjustment over the course of their first year in the program (Cornell et al., 1991a). No published studies have reported on the Rorschach performance of these subjects. Because many of these girls' Rorschach protocols were laden with unusual responses and many signs and scores conventionally assumed to indicate psychopathology, a separate analysis was undertaken to investigate the concurrent validity of their Rorschach profiles.

Rather than examine each individual Rorschach score and carry out numerous, potentially redundant correlations, we focused on three special indices of the Comprehensive System (Exner, 1993) that combine scores into empirically defensible measures of psychopathology. We chose the Schizophrenia Index to investigate whether it would discriminate disordered from creative thinking. Our rationale for using the Depression Index and CDI is based on the reports of adjustment problems and depression among some early college entrants.

Rorschach indices were compared to two standard measures of adjustment, the CPI (Gough, 1987) and the Self-Perception Profile for Adolescents (Harter, 1985). We reasoned that if the Rorschach indices were indicative of psychopathology among these girls, then they should be positively correlated with non-Rorschach adjustment measures; but on the other hand, if the Rorschach indices reflected their creative and independent thinking, the correlations should be absent or even in the opposite direction. In addition, we investigated the relation between a modified Adaptive Regression scale (Holt, 1977) and scores on the Rorschach special indices, hypothesizing that any significant correlations between the special indices and the adjustment criterion measures would be mediated by adaptive regression.

Method

Participants

The accelerant group consisted of 43 female students enrolled in an early college entrance program at a small, private, liberal arts college. Participant age ranged from 12 to 17 years *(M* = 14 years). Students were selected for this program based on a college admissions application, review of achievement scores and grades, and both parent and child interviews. The program emphasized selection of young women who seemed highly motivated, creative, and independent thinkers as well as academically qualified to accelerate their education.

The accelerants were compared to a control group of 19 girls enrolled in gifted classes at a local public high school. Although the girls in the control group were of high intelligence and had been placed in gifted programs, they lived at home and had not attempted to pursue early college entrance. Participant age ranged from 13 to 15 years *(M* = 14 years).

Students were administered the Wechsler Intelligence Scale for Children-Revised (WISC-R), the Rorschach Inkblot Test, the CPI, and the SPPA on the same day during their first month in the early college entrance program. Accelerants' full score intelligence quotients, measured by the WISC-R, ranged from 115 to 155, with a mean IQ of 133. Nonaccelerants' IQ scores ranged from 112 to 139, with a mean IQ of 124.

Rorschach Inkblot Test. Three Rorschach measures of adjustment were used, the Schizophrenia Index, the Depression Index, and the CDI, based on Exner's 1993 criteria. Two graduate students trained in the Exner system scored these three variables independently on 20 protocols and obtained intraclass correlations (Shrout & Fleiss, 1979) of .85 for the Schizophrenia Index, .79 for the Depression Index, and .90 for the CDI.

Holt's (1977) Adaptive Regression Index is derived from the defense-demand score (*DD*) and defense-effectiveness score *(DE)*. The *DD* score is based on a six-point rating scale that measures the degree to which a response demands for some defensive measure to be undertaken in order for it to adhere to socially acceptable standards. The *DE* score for each response is a rating that considers form level, defense scores, and affective expression for the response. In this study, Holt's scoring procedure was modified to simplify interpretation of results. The *DD* and *DE* scores were determined using Holt's scoring system, but to obtain the overall score reflecting adaptive regression, the sum *DE* score was subtracted from, rather than multiplied by, the *DD* score. Lower scores on this index (in which the difference between defense demand and effectiveness of defense is slight) indicate adaptive regression. Higher scores (signifying that primary process responses are not defended successfully) indicate maladjustment. Two graduate students scored the adaptive regression variable independently on 20 protocols and obtained a Pearson product-moment correlation of .85.

CPI. The CPI is a 540-item self-report instrument that includes 18 subscales and three structural scales (Gough, 1987). Because conducting analyses using this many outcome variables would increase the likelihood of obtaining some significant results by chance, three factor scales were computed: the Emotional Adjustment Index, the Social Adjustment Index, and the Autonomous Thinking Index. These CPI indices were based on equations from Nichols and Schnell's (1963) factor analytic study. In addition, the 58-item "Self-Realization" structural scale was used. Persons scoring high on the Self-Realization scale are considered to be fulfilled, optimistic, mature, able to cope with the stresses of life, and free of neurotic trends and conflicts (Gough, 1987).

SPPA. The SPPA, developed by Harter (1985), is a 36-item self-report questionnaire that contains nine subscales corresponding to different domains of self-concept. Two selected for use in this study were Social Acceptance and Global Self-Worth.

Results

T tests analyzing mean differences between accelerants and nonaccelerants on specific Rorschach indices, CPI scales, and SPPA scales are reported in Table 1.

Of the eight *t* tests conducted, there were four significant group differences. The accelerants scored higher on Schizophrenia, $t(60) = 1.71$, $p < .05$, Depression, $t(60) = 2.06$, $p < .05$, and Autonomous Thinking, $t(59) = 1.84$, $p < .05$; but lower than the nonaccelerant controls on Social Self-Concept, $t(59) = -1.95$, $p < .05$. Accelerants had significantly higher IQ scores on the WISC-R than nonaccelerants, $t(60) = 3.54$, $p < .001$. When analyses of covariance were conducted, controlling for IQ and age, the differences in these variables remained significant.

Rorschach scores were correlated with CPI and SPPA scores separately by group (see Table 2). Accelerants' Schizophrenia Index scores were correlated positively with all the outcome measures of adjustment. Accelerants who scored high on Depression showed low scores on CPI Emotional Adjustment.

Table 1
Comparisons of Accelerated and Nonaccelerated High-Ability Girls

	Accelerants[a]		Nonaccelerants[b]		
	M	*SD*	*M*	*SD*	*t test*
Rorschach index					
Schizophrenia	2.49	1.47	1.90	1.10	1.71*
Depression	4.40	1.26	3.63	1.54	2.06*
Coping deficit	2.49	1.03	2.58	1.39	-.29
CPI scale					
Emotional Adjustment	131.17	28.75	129.73	23.37	.19
Social Adjustment	81.35	13.90	82.95	14.10	-.42
Autonomous Thinking	24.03	3.84	22.01	4.25	1.84*
Self-concept scale					
General Self-worth	14.88	3.54	15.53	3.24	-.40
Social Competence	15.17	3.23	16.26	1.94	-1.95*

[a]$n = 43$. [b]$n = 19$.
*$p < .05$.

A different pattern of correlations emerged for the nonaccelerants. In contrast to the accelerants, the majority of the nonaccelerants' Rorschach indices were not related to CPI and SPPA variables. Only one correlation was significant: Nonaccelerants who scored high on Schizophrenia were lower on Autonomous Thinking. *R*-to-*z* transformations were conducted to evaluate differences in correlations between the groups. All significant group differences were for correlations involving the Schizophrenia Index.

A comparison of the number and percentage of accelerants versus nonaccelerants who scored a 4 or higher on the Schizophrenia Index was conduced by chi-square analysis. The association between group status and elevated schizophrenia scores was significant, $x^2(1, N = 62) = 3.94$, $p < .05$, with accelerants showing significantly more elevated scores than nonaccelerants. Fourteen of the 43 accelerants (34 percent) scored in the elevated range (scores of 4 or higher). In contrast, only two of the 19 nonaccelerants (10 percent) had a Schizophrenia score of 4, and none had a score of 5.

The criteria that make up the Schizophrenia Index were examined for both groups, and are presented in Table 3. Form quality was low for both groups, and both groups showed a heightened occurrence of "minus" responses outnumbering ordinary or unusual responses. In contrast, both groups showed relatively low WSS scores and only one of the 61 participants fulfilled the criteria of giving more than one Level 2 response and one or more *FABCOM* responses. The association between group status and the $M- > 1$ or $X - \% > .40$ criterion was

Table 2

Comparisons of Correlations of Rorschach Indices With and Self-Concept Subscales Among

			Rorschach
		Schizophrenia	
	Accelerants (r)[a]	**Nonaccelerants (r)**	**Group Comparison (z)**
CPI scales			
Emotional Adjustment	.31*	-.16	1.56
Social Adjustment	.34*	-.23	2.03*
Autonomous Thinking	.27*	-.42*	2.27*
Self-Realization	.38*	-.28	2.21*
Self-concept scales			
Social	.28*	-.19	1.43
General	.27*	-.27	1.79*

Note. Correlations compared by Fisher *r-to-z* transformation. [a]$n = 43$. [b]$n = 19$

significant, x^2 (1, $N = 62$) = 4.01, $p < .05$. Eighteen of the accelerants (42 percent), compared with 3 of the nonaccelerants (16 percent), met this criteria.

Further analyses were conducted to better understand the positive relations between Schizophrenia and the measures of adjustment. First, partial correlations were performed for all variables, controlling for the effects of age, IQ, and response productivity. When the effects of response productivity were partialled out of previously significant correlations between the Rorschach and adjustment variables, only the correlation between Schizophrenia and Social Adjustment became insignificant, $r(62) = .24$, $p > .05$. Controlling for age and IQ did not result in any changes in the pattern of significant correlations between Rorschach and adjustment variables.

Next, it was hypothesized that adaptive regression might help explain the relations between Rorschach and adjustment indices. The Adaptive Regression Index (AR) was significantly related to Schizophrenia, $r(62) = .44$, $p < .01$; higher scores on AR were indicative of maladjustment. To test the theory that otherwise well-adjusted participants scored high on Schizophrenia because of their creative use of adaptive regression, partial correlations were conducted, controlling for the effects of adaptive regression. These results are found in Table 4. There was no longer a significant relation between Schizophrenia and either Social Adjustment or Autonomous Thinking. The correlations between

California Psychological Inventory Composite Scales Accelerants and Nonacelerants

Indices

	Depression	
Accelerants (r)[a]	**Nonaccelerants (r)**[b]	**Group Comparison**
-.27*	-.03	-.82
.08	-.19	-.29
.18	.25	-.76
-.21	-.06	-.12
-.03	.01	.13
-.09	.04	-.13

Schizophrenia and Emotional Adjustment, Self-Realization, Social Competence, and General Self-Worth remained significant. The previously insignificant relations between Depression and both Autonomous Thinking and Self-Realization became significant when controlling for adaptive regression.

Discussion

The early college entrants in this sample scored unusually high on the Rorschach Schizophrenia Index. Fourteen of the 43 accelerants (compared to only two of the 19 nonaccelerants) scored either a 4 or a 5 on the Schizophrenia Index. These results are remarkable considering Exner's (1990) norms for 12- to 16-year-olds, in which none of his 585 participants obtained Schizophrenia scores in the elevated range.

Paradoxically, the Schizophrenia Index was positively correlated with all six outcome variables. Thus, accelerants who were high on the Rorschach index indicating thought disorder tended to be more emotionally mature, socially competent, independent and flexible, optimistic and fulfilled, and positive in their perceptions of self-worth.

The six criteria that make up the Schizophrenia Index were examined individually. Eighty-seven percent of the entire sample gave a low percentage of good form responses, suggesting that this criterion of the Schizophrenia Index should

Table 3

Frequencies of Meeting Schizophrenia Index Criteria

Criterion	**Accelerants**[a]	**Nonaccelerants**[b]
$(X+\% < .61)$ and $(S\text{-}\% < .41)$ or $(X+\% < .50)$	38 (88%)	16 (84%)
$X\text{-}\% > .29$	18 (42%)	5 (26%)
$(FQ\text{-} >= FQu)$ or $(FQ\text{-} > FQo + FQ+)$	27 (63%)	10 (53%)
(*Sum level 2 Sp. Sc.* > 1) and (*FAB2* > 0)	1 (2%)	0 (0%)
(*Raw sum of 6 Sp. Sc.* > 6) or (*Weighted sum of 6* > 17)	8 (19%)	2 (11 %)
$(M\text{-} > 1)$ or $(X\text{-}\% > .40)$	18 (42%)	3 (16%)

[a]*n* = 43. [b]*n* = 19.

be interpreted with particular caution when clinicians are assessing high-ability subjects. In contrast, just one participant gave more than one Level 2 reponse. Level 2 scores are assigned to seriously bizarre responses, and are considered to indicate severely dissociative and illogical thinking. In light of these results, this criterion, recently added to the Comprehensive System, might be particularly important to consider in judging whether Rorschach responses indicate creativity or pathology.

Accelerants also showed higher scores on the Depression Index than nonaccelerants. Elevated Depression scores reflect a preponderance of shading, achromatic color, and form dimension responses. The ability to utilize dimension, depth, and shading in perceiving the cards might also be an indication of creativity. Alternatively, the elevated Depression scores could signify higher levels of depression in accelerated students. Cornell et al. (1991b), in a study of socioemotional adjustment in this same sample of early college entrants, found that over half were reported by staff as experiencing noteworthy depression. In addition, the accelerants showed lower scores on Social Self-Concept than did the nonaccelerants. These results are consistent with Lupkowski et al.'s (1992) findings that early college entrants experienced a significant drop in self-esteem upon entrance to college.

The Depression Index was negatively related to the CPI emotional adjustment index, supporting the more conventional hypothesis that indicators of maladjustment on the Rorschach would be related to other indicators of maladjustment. Accelerants who were more depressed according to the Rorschach also showed signs of less healthy emotional adjustment according to the CPI.

Table 4
Partial Correlations of Rorschach Indices With Adjustment Measures

	Rorschach Indices			
	Schizophrenia		Depression	
	Accelerants[a]	Nonaccelerants[b]	Accelerants[a]	Nonaccelerants[b]
CPI scales				
Emotional Adjustment	.30*	-.15	-.35	-.02
Social Adjustment	.18	-.22	.00	-.16
Autonomous Thinking	.14	-.42*	-.27*	.25
Self-Realization	.32*	-.28	-.31*	-.06
Self-concept scales				
Social	.29*	-.19	-.03	.03
General	.27*	-.27	-.12	.04

[a]$n = 43$. [b]$n = 19$.
*$p < .05$.

High-Ability Adolescents' Use of Adaptive Regression

When accelerants high on adaptive regression defend their unusual, primary process responses logically, adaptively, and in a socially acceptable manner, the shared variance between Schizophrenia and both Social Adjustment and Autonomous Thinking can be attributed to adaptive regression. However, Schizophrenia is still correlated positively with several other outcome variables of adjustment. These results offer only partial support for the hypothesis that a creative thinking style accounts for the correlations between Schizophrenia and adjustment.

When Holt's (1970) adaptive regression system was applied, several of the seemingly disordered subjects, according to the Rorschach, gave responses that were logically defended in an adaptive style and reflected their creative and imaginative orientation to the task. One high-ability subject, who scored high on Exner's Schizophrenia Index, gave a response to Card VIII that contained primary process unusual form, and several special scores, yet reflected her creative thinking style: "A demon in the shape of a boar, rising up into a forest from Hades' Underground taking two souls with him." According to Holt (1970), reference to cultural context (e.g., "Hades' Underground") is one way subjects successfully defend primary process in their responses.

Some high-ability subjects gave responses that would qualify as Special Scores according to Exner's system, but did not necessarily indicate disordered thinking. For example, one subject's *FABCOM* response of "little bugs having a party" can be interpreted as creative and imaginative, rather than a break in reality testing. Thus, the Exner Schizophrenia Index, and its component, must be examined closely to determine if they reflect psychopathology or adaptive regression in high-ability adolescents.

In conclusion, the Schizophrenia Index must be interpreted with caution when applied to early college entrants and perhaps to other groups of highly capable but unconventional subjects, Such subjects may be motivated to produce original or unusual responses. As one accelerant remarked, "I could say a bat, but everyone sees a bat, so I won't."

Acknowledgements

Data collection for this project was supported by grants from the Appalachia Research Laboratory.

We thank the students and staff who participated in this study.

This study replicated and extended previous work by Houlihan (1989).

References

Archer, R. R., & Gordon, R. A. (1988). MMPI and Rorschach indices of schizophrenic and depressive diagnoses among adolescent inpatients. *Journal of Personality Assessment 52,* 276-287.

Barrett H. O. (1957). An intensive study of 32 gifted children. *Personnel and Guidance Journal, 36,* 192-194.

Brody, L. B., & Benbow, C. P. (1987). Accelerative strategies: How effective are they for the gifted? *Gifted Child Quarterly, 31,* 105-110.

Cornell, D. G., Callahan, C., & Loyd, B. (1991a). Personality growth of female early college entrants: A controlled, prospective study. *Gifted Child Quarterly 35,* 13-43.

Cornell, D. G., Callahan, C., & Loyd, B. (1991a). Socioemotional adjustment of adolescent girls enrolled in a residential acceleration program. *Gifted Child Quarterly 35,* 58-66.

Dudek S. Z. (1970). Effects of different types of psychotherapy on the personality as a whole. *Journal of Nervous and Mental Disease, 150,* 329-345.

Dudek, S. Z., & Cumberland-Bouhadana, G. (1982). Primary process in creative persons. *Journal of Personality Assessment,* 46, 239-247.

Exner, J. E. Jr. (1990). *A Rorschach workbook for the comprehensive system* (3rd ed,). Asheville NC: Rorschach Workshops

Exner, J. E., Jr. (1993). *The Rorschach A comprehensive system: Vol. I Basic foundations* (3rd ed.). New York Wiley.

Gallagher, J.J., & Crowder, T. (1957). The adjustment of gifted children in the regular classroom. *Exceptional Children, 23,* 306-319.

Gallucci, N. T. (1989), Personality assessment with children of superior intelligence: Divergence versus psychopathology. *Journal of Personality Assessment, 53,* 749-760.

Gough, H. G. (1987). *California Psychological Inventory manual* (Rev. ed.). Palo Alto, CA: Consulting Psychologists Press.

Harter, S. (1985). *Manual for the Self-Perception Profile for Adolescents.* Denver, CO: University of Denver.

Holt, R. R. (1970). *Manual for the scoring of primary process manifestations in Rorschach responses.* New York: Research Center for Mental Health, New York University.

Holt, R. R. (1977). A method for assessing primary process manifestations and their control in Rorschach responses. In M. A. Rickers-Ovsiankina (Ed.), *Rorschach psychology* (pp. 375-420). Huntington, NY: Krieger.

Houlihan, T. (1989). *Clarifying the use of the Rorschach with high ability adolescent females.* Unpublished doctoral dissertation, University of Virginia.

Janos, P.M., Robinson, N. M., Carter, C., Chapel, A., Cufley, R., Curland, M., Daily, M., Gailland, M., Heinzig, M., Kehl, H., Lu, S., Sherry, D., Stoloff, J., & Wise, A. (1988). A cross-sectional developmental study of the social relations of students who enter college early. *Gifted Child Quarterly,* 32, 210-245.

Kris, E. (1952). *Psychoanalytic explorations in art.* New York: International Universities Press.

Lupkowski, A., Whitmore, M., & Ramsay, A. (1992). The impact of early entrance to college on self-esteem: A preliminary study. *Gifted Child Quarterly,* 36, 87-90.

Netter, B., & Viglione, D., Jr. (1994). An empirical study of malingering schizophrenia on the Rorschach. *Journal of Personality Assessment,* 62, 45-57.

Nichols, R. C., & Schnell, R. R. (1963). Factor scales for the California Psychological Inventory. *Journal of Consulting Psychology,* 27, 228-235.

Pine, F., & Holt, R. R. (1960). Creativity and primary process: A study of adaptive regression. *Journal of Abnormal and Social Psychology,* 61, 370-379.

Robinson, N., & Janos, P. (1986). Psychological adjustment in a college-level program of marked acceleration. *Journal of Youth and Adolescence,* 15, 51-60.

Selig, K. (1958). *Personality structure as revealed by the Rorschach technique of a group of children who test at or above 170 I.Q. on the 1937 revision of the Stanford-Binet scale.* Unpublished doctoral dissertation, New York University.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rating reliability. *Psychological Bulletin, 86,* 420-428.

Sternberg, R. J., & Davidson, J. E. (1983). Insight in the gifted. *Educational Psychologist, 18,* 51-57.

BEST COPY AVAILABLE

The Mensa Awards for Excellence in Research

Mensa Awards for Excellence in Research are given each year to eight to ten scientists who have published outstanding research papers in peer-reviewed professional journals. This worldwide competition is sponsored jointly by the Mensa Education and Research Foundation and Mensa International, Ltd.

Typically, half the awards are given to established, senior scientists and half to researchers who have, in the past five years, entered into research into the nature of human intelligence or giftedness, education for the intellectually gifted, etc. Eligible fields of research have included psychology, education, sociology, neurology, physiology, biochemistry, and psychometrics.

Each award consists of $500 and a certificate. Many of the winning articles are reprinted in the *Mensa Research Journal.*

Judging is done by the joint American Mensa, Ltd./Mensa Education and Research Foundation Research Review Committee.

For additional information about how to enter a paper into this competition, write to MERF, Awards for Excellence in Research, 1229 Corporate Drive West, Arlington, TX 76006, USA.

MENSA 45
Research Journal

Mensa Education and Research Foundation
1229 Corporate Drive West
Arlington, TX 76006

Address service requested. Return postage guaranteed

NON-PROFIT
ORGANIZATION
U.S. POSTAGE
PAID
Arlington, TX
Permit No. 922

***ALL FOR ADC 220
G#100058523 G
ERIC CLEARINGHOUSE
1920 ASSOCIATION DR
RESTON VA 20191-1545